GENERAL APPROACH TO THE FLUCTUATIONS PROBLEM IN RANDOM SEQUENCE COMPARISON

Jüri Lember*, Heinrich Matzinger, Felipe Torres†

November 22, 2012

Jüri Lember, Tartu University, Institute of Mathematical Statistics
Liivi 2-513 50409, Tartu, Estonia. *E-mail:* jyril@ut.ee
Heinrich Matzinger, Georgia Tech, School of Mathematics
Atlanta, Georgia 30332-0160, U.S.A. *E-mail:* matzing@math.gatech.edu
Felipe Torres, Münster University, Institute for Mathematical Statistics
Einsteinstrasse 62, 48149 - Münster, Germany. *E-mail:* ftorrestapia@math.uni-muenster.de

Abstract

We present a general approach to the problem of determining the asymptotic order of the variance of the optimal score between two independent random sequences defined over an arbitrary finite alphabet. Our general approach is based on identifying random variables driving the fluctuations of the optimal score and conveniently choosing functions of them which exhibit certain monotonicity properties. We show how our general approach establishes a common theoretical background for the techniques used by Matzinger *et al* in a series of previos articles [6, 8, 20, 24, 26, 37] studying the same problem in especial cases. Additionally, we explicitely apply our general approach to study the fluctuations of the optimal score between two random sequences over a finite alphabet (closing the study as initiated in [26]) and of the length of the longest common subsequences between two random sequences with a certain block structure (generalizing part of [37]).

Keywords. *Random sequence comparison, longest common sequence, fluctuations, Waterman conjecture.*

AMS. 60K35, 41A25, 60C05

*Supported by the Estonian Science Foundation Grant nr. 9288 and targeted financing project SF0180015s12

†corresponding author and research supported by the DFG through the SFB 878 at University of Münster

1 Introduction

1.1 Sequence comparison setting

Throughout this paper $X = (X_1, X_2, \dots X_n)$ and $Y = (Y_1, Y_2, \dots Y_n)$ are two random strings, usually referred as sequences, so that every random variable X_i and Y_i take values on a finite alphabet $\mathbb{A}$. We shall assume that the sequences X and Y have the same distribution and are independent. The sample space of X and Y will be denoted by $\mathcal{X}_n$. Clearly $\mathcal{X}_n \subseteq \mathbb{A}^n$ but, depending on the model, the inclusion can be strict.
The problem of measuring the similarity of X and Y is central in many areas of applications including computational molecular biology [9, 15, 34, 35, 41] and computational linguistics [42, 27, 31, 32]. In this paper, we adopt the same notation as in [25], namely we consider a general scoring scheme, where $S : \mathbb{A} \times \mathbb{A} \to \mathbb{R}^+$ is a *pairwise scoring function* that assigns a score to each couple of letters from $\mathbb{A}$. An *alignment* is a pair (ρ, τ) where $\rho = (\rho_1, \rho_2, \dots, \rho_k)$ and $\mu = (\tau_1, \tau_2, \dots, \tau_k)$ are two increasing sequences of natural numbers, i.e. $1 \le \rho_1 < \rho_2 < ... < \rho_k \le n$ and $1 \le \tau_1 < \tau_2 < \dots < \tau_k \le n$. The integer k is the number of aligned letters, $n - k$ is the number of *gaps* in the alignment. Note that our definition of gap slightly differs from the one that is commonly used in the sequence alignment literature, where a gap consists of maximal number of consecutive *indels* (insertion and deletion) in one side. Our gap actually corresponds to a pair of indels, one in X-side and another in Y-side. Since we consider the sequences of equal length, to every indel in X-side corresponds an indel in Y-side, so considering them pairwise is justified. In other words, the number of gaps in our sense is the number of indels in one sequence. We also consider a *gap price* δ. Given the pairwise scoring function S and the gap price δ, the score of the alignment (π, μ) when aligning X and Y is defined by

$$U_{(\rho,\tau)}(X, Y) := \sum_{i=1}^{k} S(X_{\rho_i}, Y_{\tau_i}) + \delta(n - k).$$

In our general scoring scheme δ can also be positive, although usually $\delta \le 0$ penalizing the mismatch. For negative δ, the quantity $-\delta$ is usually called the *gap penalty*. The optimal alignment score of X and Y is defined to be

$$L_n := L(X, Y) := \max_{(\pi,\mu)} U_{(\rho,\tau)}(X, Y), \tag{1.1}$$

where the maximum above is taken over all possible alignments. To simplify the notation, in what follows, we shall denote $Z := (X, Y)$ so that $L_n = L(Z)$.

When $\delta = 0$ and the scoring function assigns one to every pair of similar letters and zero to all other pairs, i.e.

$$S(a, b) = \begin{cases} 1, & \text{if } a = b; \\ 0, & \text{if } a \neq b. \end{cases} \tag{1.2}$$

then $L(Z)$ is just the maximal number of aligned letters – the length of the *longest common subsequence (abbreviated by LCS)*. The longest common subsequence is probably the most common measure of global similarity between strings.

1.2 The variance problem: history and the state of art

Since X, Y are random string, the optimal score L_n is a random variable. In order to distinguish related pairs of strings from unrelated ones, it is relevant to study the distribution of L_n for independent sequences. When X and Y are take from an ergodic processes then, by Kingman's subbaditive ergodic theorem, there exists a constant γ such that

$$\frac{L_n}{n} \to \gamma \text{ a.s. and in } L_1, \text{ as } n \to \infty. \tag{1.3}$$

In the case of LCS, namely when S is taken as in (1.2), the constant γ is sometimes called the *Chvatal-Sankoff constant* and its value, although well estimated (see [3, 7, 5, 36, 13, 12, 33, 28, 23, 19, 22]) remains unknown even for as simple cases as i.i.d. Bernoulli sequences. The existence of γ was first noticed by Chvatal and Sankoff in their pioneering paper [10], where they proved that the limit

$$\gamma := \lim_{n\to\infty} \frac{EL_n}{n} \tag{1.4}$$

exists. In [1] the rate of the convergence in (1.4) was for the first time established, and in [25] the authors improved the previous results introducing a new technique based on entropy and combinatorics, which gives a little more bout the path structure of the optimal alignments.

The fluctuations of L_n. To make inferences on L_n, besides the convergence (1.3), the size of the variance $\mathrm{Var}[L_n]$ is essential. Unfortunately, not much is known about $\mathrm{Var}[L_n]$ and its asymptotic order is one of the central open problems in string matching theory. Monte-Carlo simulations lead Chvatal and Sankoff in [10] to conjecture that, in the case of LCS, $\mathrm{Var}[L_n] = o(n^{\frac{2}{3}})$ for i.i.d. $\frac{1}{2}$-Bernoulli sequences. Using an Efron-Stein type of inequality, Steele [36] proved that there exist a constant $B < \infty$ such that $\mathrm{Var}[L_n] \leq Bn$. In [38], and always in the LCS case, Waterman asks whether this linear bound can be improved. His simulations show that this is not the case and $\mathrm{Var}[L_n]$ should grow linearly. Still in the LCS case, Boutet de Monvel [7] interprets his simulations the same way. In a series of papers containing different settings, Matzinger *et al.* have been investigating the asymptotic behavior of $\mathrm{Var}[L_n]$. Their goal is to find out whether there exists a constant $b > 0$ (not depending on n) such that $\mathrm{Var}[L_n] \geq bn$. Together with Steele's bound, this means that $bn \leq \mathrm{Var}[L_n] \leq Bn$, i.e. $\mathrm{Var}[L_n] = \Theta(n)$ (we say that a sequence a_n is of order $\Theta(n)$ if, there exist constants $0 < b < B < \infty$ such that $bn \leq a_n \leq Bn$ for all n large enough). So far, most of the research to show that $\mathrm{Var}[L_n] = \Theta(n)$ has been done in the case of LCS:

- In [8], X is a $\frac{1}{2}$-Bernoulli binary sequence and Y is a non-random periodic binary sequence,
- In [6], X is a $\frac{1}{2}$-Bernoulli binary sequence and Y is an i.i.d. random sequence over a 3 – symbols alphabet,

- In [20], both X and Y are $\frac{1}{2}$-Bernoulli binary sequences but they are aligned by using a score function which gives more weight when aligning ones than aligning zeros,
- In [24], both X and Y are i.i.d. binary sequences, but one symbol has much smaller probability than the other. That is a so called *case of low entropy*.
- In [37], both X and Y are binary sequence having a multinomial block structure. That is, for the first time, a so called *case of high entropy*.

Another related string matching problem is the so called *longest increasing subsequence (abbreviated by LIS) problem*. Given a sequence X, to find the LIS of X is to find an increasing sequence of natural numbers $1 \leq i_1 \leq i_2 \leq \cdots \leq i_k \leq n$ such that $X_{i_1} \leq X_{i_2} \leq \cdots \leq X_{i_k}$. The LIS problem can be seen as a particular case of the LCS problem, in the following way: Let $X := 1\,2\cdots n$ be the sequence of the first n increasing integers and let $\sigma(X) := \sigma(1)\sigma(2)\cdots\sigma(n)$ be the sequence of its random permutation. Then, a LIS of X corresponds to a LCS of X and $\sigma(X)$. Due to this equivalence, it was thought that the LIS and the LCS have fluctuations of the same order, which now we know it is not true. In this direction Houdre, Lember and Matzinger [21] studied an hydrid problem, namely the fluctuations of ℓ_n defined as the length of the longest common increasing subsequence of X and Y, where X and Y are i.i.d. $\frac{1}{2}$-Bernoulli binary sequences, and a longest common increasing subsequence of X and Y is just a LIS of X and of Y simultaneously. They showed that $n^{-1/2}(\ell_n - E\ell_n)$ converges in law to a functional of two Brownian motions, which implies that $\mathrm{Var}[\ell_n] = \Theta(n)$ holds as well. There is also a connection between the LCS of two random sequences and a certain passage percolation problem [1].
For the case of general scoring, to our best knowlodge, the only previous partial results on fluctuations are cointained in [26].

1.3 Main results and the organization of the paper

Recall that we aim to prove (or disprove) the order of the variance $\mathrm{Var}[L_n] = \Theta(n)$ and due to Steele's upper bound it suffices to prove (or disprove) the existence of $b > 0$ so that $\mathrm{Var}[L_n] \geq bn$. All available proofs (by Matzinger *et al.*) of the existence of such b follow more or less the same philosophy and can be split into two parts, strategy that we call *two-step approach*. The first part of this approach is to find a random mapping independent of Z, usually called a *random transformation*,

$$\mathcal{R} : \mathcal{X}_n \times \mathcal{X}_n \to \mathcal{X}_n \times \mathcal{X}_n$$

such that, for most of the outcomes $z \in \mathcal{X}_n \times \mathcal{X}_n$ of Z, increases the score at least by some fixed amount $\epsilon_o > 0$. More precisely, the random transformation should be that for some $\alpha > 0$ there exists a set $B_n \subset \mathcal{X}_n \times \mathcal{X}_n$ having probability at least $1 - \exp[-n^\alpha]$ so that, for every $z \in B_n$, the expected score of $\mathcal{R}(z)$ exceeds the score of z by ϵ_o (where the expectation is taken over the randomness involved in the transformation), namely

$$E\big[L(\mathcal{R}(z))\big] \geq L(z) + \epsilon_o.$$

Before stating this requirement formally, let us introduce a useful notation: $\tilde{Z} := \mathcal{R}(Z)$. Thus $\tilde{Z}$ is obtained from Z by applying a random modification to Z and the additional randomness is independent of Z. Formally, the first step of the approach is to find a random transformation so that for some universal constants $\alpha > 0$ and $\epsilon_o > 0$, the following inequality holds:

$$P\big(E[L(\tilde{Z}) - L(Z)|Z] \geq \epsilon_o\big) \geq 1 - \exp[-n^\alpha]. \tag{1.5}$$

Besides (1.5), the random transformation has to satisfy some other requirements. This other requirements and their influence on the fluctuations of L_n form the second step of the two-step approach. Roughly speaking, there should also exist an associated function of Z, let us call it $\mathfrak{u}(Z)$, so that applying the random transformation to Z increases the value of $\mathfrak{u}$. The variance of L_n can be then lower bounded by the variance of $U := \mathfrak{u}(Z)$ so that the constant b exists if the variance of $\mathfrak{u}(Z)$ is linear on n. This second step is formally presented and explained with details in Subsection 2.3, where we also briefly introduce the random transformation and the random variable U used so far. Let us remark that in earlier articles of the subject [6, 20, 24, 37], the random transformation is not explicitly defined, but the variance driving random variable U is always there, and one can easily define the random transformation as well. Let us also mention that to show (1.5) for a suitable chosen transformation is not an easy task and, typically, needs a lot of effort. The second step of the approach consists of showing that (1.5) implies the existence of b. This proof depends on the model, on the chosen transformation and its components (vectors U and V, see Subsection 2.3). One of the goals of the present paper is to present a general setup and a general proof for the second step. With such a general proof in hand, all the the future proofs of the existence of b could be remarkably shortened and simplified. Our general approach is based on the same strategy as in [24, 37], but remarkably shorter and simpler (see also Remark 6 before the proof of Theorem 2.2). These general results are Theorem 2.1 and Theorem 2.2, both presented in Section 2.
In order to see in action our two-step approach, we include two applications which bring us up two new fluctuation results:

First application: optimal score of i.i.d. sequences. In Section 3, X and Y are $\mathbb{A}$-valued i.i.d. random variables, being compared under the general scoring scheme described in Subsection 1.1. In this case, the random transformation consists of uniformly choosing a specific letter $a \in \mathbb{A}$ and turning it into another specific letter $b \in \mathbb{A}$. In [26], it has been proven that when the gap price is relatively low and the scoring function S satisfies some mild asymmetry assumptions, then the described transformation satisfies (1.5), see Theorem 3.1 and the remarks after it. Thus, for sufficiently low gap price δ, the first step holds true. In Section 3, we show that all other assumptions of Theorem 2.1 and Theorem 2.2 are fulfilled, so that the second step holds true and thus there exists the desired constant b (Theorem 3.2). Hence, Section 3 completes the study started in [26] and, to our best knowledge, we obtain the first result where the order of variance $\mathrm{Var}[L_n] = \Theta(n)$ is proven in a setup other than LCS of binary sequences. It is important to note that for the second step (Theorem 3.2), no assumption on δ nor on the scoring

function S are needed. Hence, whenever the assumptions in the first step can be relaxed (i.e. generalization of Theorem 3.1), the second step still holds true and the order of variance $\mathrm{Var}[L_n] = \Theta(n)$ can be automatically deduced.

Second application: The length of the LCS of random i.i.d. block sequences. Unfortunately, the current assumption on the gap price δ makes Theorem 3.1 not applicable to the length of the LCS of two independent i.i.d. sequences, thus this case (except the special model in [24]) is still open. In order to approach this still open question from another point of view, in Section 4 we consider X and Y not to be i.i.d. sequences any more, but we keep $\mathbb{A} = \{0, 1\}$ consisting of two colours (i.e. the sequences are still binary) and the scoring function to be (1.2), also $\delta = 0$. Hence, L_n is the length of the LCS. The difference from the setup considered in Section 3 lies in the random structure of X and Y. Let us briefly explain the model. Note that any binary sequence can be considered as a concatenation of *blocks* with switching colors (from 0 to 1 or viceversa). Here a block is merely a subword of the sequence having all letters of the same color and a different color before and after it. Hence, every binary sequence is fully determined by the lengths of its blocks and the color of its first block. Therefore, every infinite i.i.d. Bernoulli sequence X with parameter $\frac{1}{2}$ can be considered as an i.i.d sequence of blocks whose lengths are geometrically distributed, where the first block has colour either 0 or 1 with probability $\frac{1}{2}$. X can be, in a sense, approximated by a (binary) random sequence $\hat{X}$ with finite range of possible block lengths. Indeed, the probability of finding a very long block in X is very small, hence such an approximation of X by $\hat{X}$ is justified (note that although the blocks remain to be i.i.d, $\hat{X}$ is not an i.i.d. sequence any more). This is the situation in Section 4: instead of considering X (and Y) as the first n elements of an i.i.d. infinite Bernoulli sequence with parameter $\frac{1}{2}$, we take them as the first n elements of an infinite sequence $X_1, X_2, \ldots$ obtained by i.i.d. concatenating blocks of alternated colours of random lengths distributed on $\{l-1, l, l+1\}$, where $l > 2$ is a fixed integer. For a formal description of this block model see Subsection 4.1. The restriction that the block lengths can only have three possible values is made in order to have a simplified exposition of the technique. We believe that the results in Section 4 also hold for any finite range of possible block lengths.

Considering such a block model is motivated by the following arguments. First, it is a common practice in random sequence comparison to approximate a target model (i.i.d. Bernoulli sequences, in our case) by some more tractable model. In random sequence comparison, the more tractable model typically has lower entropy. Secondly, as it is shown in [37], for the case where all three possible block lengths have equal probability, there exists a random transformation so that (1.5) holds, see Lemma 4.6. The random transformation in this case – let us call it the *block-transformation* – is the following: pick uniformly an arbitrary block of X with length $l-1$ and independently an arbitrary block of X with length $l+1$. Then, change them both into blocks of length l. Thus, in this particular case where all block lengths have equal probability, the first-step is accomplished. In Section 4, we show that the block-transformation and corresponding random variables satisfy all other assumptions of Theorem 2.1 and Theorem 2.2, so that the existence of b follows

(Theorem 4.1). Thus, for the case of equiprobable block lengths, the order of variance $\mathrm{Var}[L_n] = \Theta(n)$ has now been proved. Since the uniform distribution of of block lengths was not used in the second step (see Theorem 4.1), it follows that the same order of variance automatically holds if an equivalent to Lemma 4.6 without the uniform distribution assumption can be shown. Again, we believe that such a generalization is true.

2 The two-step approach

2.1 Preliminaries

Proposition 2.1 *Let N be an integer-valued random variable taking values on interval I. Let $f : I \to \mathbb{R}$ be a monotone function so that for a $c > 0$, $f(k) - f(k-1) \geq c$ (or $f(k-1) - f(k) \geq c$) for every $k, k-1 \in I$. Then* $\mathrm{Var}[f(N)] \geq c^2 \mathrm{Var}[N]$.

For a proof of this statement see [6]. The next corollary replaces the more involved Lemma 3.3 in [6] or Lemma 5.0.3 in [37]. In our general approach, we need a simpler version because we use the decomposition (2.9) (see Remark 6. after Theorem 2.1):

Corollary 2.1 *Let N be an integer-valued random variable taking values on the set $\mathcal{Z} := \{z_1, z_2, \ldots\} \subset \mathbb{Z}$, . Let $k_o := \sup_{i \geq 2}(z_i - z_{i-1}) < \infty$. Let f be an increasing function defined on $\mathcal{Z}$ so that for $\delta > 0$ it holds*

$$f(z_i) - f(z_{i-1}) \geq \delta, \quad \forall i \geq 2.$$

Then

$$\mathrm{Var}[f(N)] \geq \frac{\delta^2}{k_o^2} \mathrm{Var}[N].$$

Proof. Let M be a random variable taking values on the set $I = \{1, 2, \ldots\}$ defined as follows: $M = i$ iff $N = z_i$. Let g be an increasing function on I defined as follows: $g(i) := f(z_i)$. Thus $g(i+1) - g(i) \geq \delta$ and $g(M) = f(N)$. By Proposition 2.1

$$\mathrm{Var}[f(N)] = \mathrm{Var}[g(M)] \geq \delta^2 \mathrm{Var}[M] \geq \big(\frac{\delta}{k_o}\big)^2 \mathrm{Var}[N],$$

where the last inequality follows from the inequality $\mathrm{Var}[N] \leq k_o^2 \mathrm{Var}[M]$.

■

Lemma 2.1 (Chebychev's inequality) *Let U be a random variable, then for any constant $\zeta > 0$ we have*

$$P\Big(\; |U - \mathrm{E}[U]| \geq \zeta \sqrt{\mathrm{VAR}[U]} \; \Big) \leq \frac{1}{\zeta^2}. \tag{2.1}$$

Lemma 2.2 (Höffding's inequality) *Let $a > 0$ be constant and $V_1, V_2, \ldots$ be an i.i.d sequence of bounded random variables such that:*

$$\mathrm{P}(|V_i - \mathrm{E}[V_i]| \leq a) = 1$$

for every $i = 1, 2, \ldots$ Then for every $\Delta > 0$, we have that:

$$\mathrm{P}\left(\left|\frac{V_1 + \cdots + V_n}{n} - \mathrm{E}[V_1]\right| \geq \Delta\right) \leq 2\exp\left(-\frac{\Delta^2}{2a^2} \cdot n\right) \tag{2.2}$$

The following lemma follows from the local central limit theorem (section 2.5 in [18]):

Lemma 2.3 *Let $X \sim B(m, p)$ be a binomial random variable with parameters m and p. Then, for any constant $\beta > 0$, there exists $b(\beta)$ and $m_0(\beta)$ so that for every $m > m_o$ and*

$$i \in [mp - \beta\sqrt{m}, mp + \beta\sqrt{m}] =: I_m,$$

it holds

$$P(X = i) = \binom{m}{i} p^i (1-p)^{m-i} \geq \frac{1}{b\sqrt{m}}. \tag{2.3}$$

Moreover, there exists an universal constant $c_1(\beta) > 0$ and $m_1(\beta)$ so that for every $m > m_1$

$$\mathrm{Var}[X | X \in I_m] \geq c_1 m. \tag{2.4}$$

Applying Lemma 2.3 repeatedly on marginals, we obtain a multinomial corollary:

Corollary 2.2 *Let (X, Y, Z) be a multinomial random vector with parameters m and p_1, p_2, p_3 such that $p_1 + p_2 + p_3 = 1$. Then, for any constant $\beta > 0$, there exists $b(\beta)$ and $m_0(\beta)$ so that for every $m > m_o$ and*

$$(i, j) \in [mp_2 - \beta\sqrt{m}, mp_2 + \beta\sqrt{m}] \times [mp_1 - \beta\sqrt{m}, mp_1 + \beta\sqrt{m}],$$

it holds

$$P(X = i, Y = j) = \binom{m}{i\ j\ (m-i-j)} p_1^i p_2^j p_3^{m-i-j} \geq \frac{1}{bm}. \tag{2.5}$$

2.2 General fluctuations results

Let $\mathcal{X}_n$ be the sample space of X and Y so that $\mathcal{X}_n \times \mathcal{X}_n$ is the sample space of $Z := (X, Y)$. In the following, we are considering the functions

$$\mathfrak{u}: \quad \mathcal{X}_n \times \mathcal{X}_n \to \mathbb{Z}, \quad \mathfrak{v}: \quad \mathcal{X}_n \times \mathcal{X}_n \to \mathbb{Z}^d$$

so that $U := \mathfrak{u}(Z)$ (resp. $V := \mathfrak{v}(Z)$) is an integer values random variable (resp. vector). We shall denote by $\mathcal{S}_n$, $\mathcal{S}_n^U$ and $\mathcal{S}_n^V$ the support of (U, V), U and V, respectively. Hence $\mathcal{S}_n \subset \mathbb{Z}^{d+1}$, $\mathcal{S}_n^U \subset \mathbb{Z}$ and $\mathcal{S}_n^V \subset \mathbb{Z}^d$. For every $v \in \mathcal{S}_n^V$, we define the fiber of $\mathcal{S}_n^U$ as follows

$$\mathcal{S}_n(v) := \{u \in \mathcal{S}_n^U : (u, v) \in \mathcal{S}_n\}.$$

For any $(u,v) \in \mathcal{S}_n$, let

$$l(u,v) := E[L(Z)|U=u, V=v].$$

In what follows, we shall often consider the conditional distribution of U given that V takes a particular value v. Therefore, we shall denote by $U_{(v)}$ a random variable that has this conditional distribution, i.e. for any $z \in \mathbb{Z}$, it holds

$$P(U_{(v)} = z) = P(U = z|V = v).$$

We shall also consider the sets of "typical values" of V and $U_{(v)}$. More precisely, we shall define the sets $\mathcal{V}_n \subset \mathcal{S}_n^V$ that contain (in some sense) the values of V that are of our interest. Similarly, for every $v \in \mathcal{V}_n$, we shall define the sets $\mathcal{U}_n(v)$ that (again in some sense) contains the values of $U_{(v)}$ that are of our interest. Roughly speaking, in what follows we shall always condition on the events $\{V \in \mathcal{V}_n\}$ and $\{U_{(v)} \in \mathcal{U}_n(v)\}$.

Theorem 2.1 *Assume the existence of sets $\mathcal{V}_n \subset \mathcal{S}_n^V$ and $\mathcal{U}_n(v) \subset \mathcal{S}_n(v)$, for $v \in \mathcal{V}_n$, so that for some constants $\delta > 0$ and $k_o \in \mathbb{N}$, the following conditions hold:*

1) *For every $v \in \mathcal{V}_n$ and $u_1, u_2 \in \mathcal{U}_n(v)$ such that $u_1 < u_2$, it holds*

$$l(u_2, v) - l(u_1, v) \geq \delta. \tag{2.6}$$

2) *There exists ψ_n so that for every $v \in \mathcal{V}_n$, the following lower bound holds*

$$\mathrm{Var}[U_{(v)}|U_{(v)} \in \mathcal{U}_n(v)] \geq \psi_n. \tag{2.7}$$

3) *There exists $k_o > \infty$ so that for every $v \in \mathcal{V}_n$ and $u_1 \in \mathcal{U}_n(v)$, there exists an $u_2 \in \mathcal{U}_n(v)$ so that $|u_1 - u_2| \leq u_1 + k_o$.*

Then

$$\mathrm{Var}[L(Z)] \geq \left(\frac{\delta}{k_o}\right)^2 \cdot \psi_n \cdot \sum_{v \in \mathcal{V}_n} P(U_{(v)} \in \mathcal{U}_n(v)) P(V = v). \tag{2.8}$$

Proof. It is clear that

$$\mathrm{Var}[L(Z)] = E\big(\mathrm{Var}[L(Z)|U,V]\big) + \mathrm{Var}\big(E[L(Z)|U,V]\big) \geq \mathrm{Var}\big(l(U,V)\big). \tag{2.9}$$

We aim to bound $\mathrm{Var}\big(l(U,V)\big)$ from below. We condition on V and use the same formula to get

$$\begin{aligned}
\mathrm{Var}\big(l(U,V)\big) &= E\big(\mathrm{Var}[l(U,V)|V]\big) + \mathrm{Var}\big(E[l(U,V)|V]\big) \geq E\big(\mathrm{Var}[l(U,V)|V]\big) \\
&= \sum_{v \in \mathcal{S}_n^V} \mathrm{Var}[l(U,v)|V=v] P(V=v) \geq \sum_{v \in \mathcal{V}_n} \mathrm{Var}[l(U,v)|V=v] P(V=v) \\
&= \sum_{v \in \mathcal{V}_n} \mathrm{Var}[l(U_{(v)},v)] P(V=v).
\end{aligned} \tag{2.10}$$

Conditioning on the event $\{U_{(v)} \in \mathcal{U}_n\}$, we see that

$$\mathrm{Var}[l(U_{(v)}, v)] \geq \mathrm{Var}[l(U_{(v)}, v)|U_{(v)} \in \mathcal{U}_n(v)]P(U_{(v)} \in \mathcal{U}_n).$$

By assumption **1)**, on the set $\mathcal{U}_n(v)$ the function l satisfies (2.6). By assumption **3)**, the two consecutive atoms of $\mathcal{S}_n(v) \cap \mathcal{U}_n$ are at most k_o apart from each other. By Corollary 2.1, thus

$$\mathrm{Var}[l(U_{(v)}, v)|U_{(v)} \in \mathcal{U}_n] \geq \frac{\delta^2}{k_o^2} \cdot \mathrm{Var}[U_{(v)}|U_{(v)} \in \mathcal{U}_n]. \tag{2.11}$$

Thus (2.10) can lower bounded by

$$\begin{aligned}\sum_{v \in \mathcal{V}_n} \mathrm{Var}[l(U_{(v)}, v)]P(V = v) &\geq \left(\frac{\delta}{k_o}\right)^2 \cdot \sum_{v \in \mathcal{V}_n} \mathrm{Var}[U_{(v)}|U_{(v)} \in \mathcal{U}_n(v)]P\big(U_{(v)} \in \mathcal{U}_n(v)\big)P(V = v) \\ &\geq \left(\frac{\delta}{k_o}\right)^2 \cdot \psi_n \cdot \sum_{v \in \mathcal{V}_n} P\big(U_{(v)} \in \mathcal{U}_n(v)\big)P(V = v).\end{aligned}$$

■

Remarks:

1. The theorem above is non-asymptotic. It means that n is fixed and, therefore, could be removed from the statement. However, writing the theorem in such a way, we try to stress out that δ and k_o should be independent of n when applying the theorem. Obviously X, Y, Z, U, V will depend on n too, but we do not explicitely include that in the notation.
2. In order to get a linear lower bound from (2.8), it suffices to show that for some constant $b > 0$ it holds,

   $$\psi_n \cdot \sum_{v \in \mathcal{V}_n} P(U_{(v)} \in \mathcal{U}_n(v))P(V = v) \geq b\, n$$

   Typically ψ_n is linear on n so that for a constant $d > 0$ we will have $\psi_n \geq dn$, and the sets $\mathcal{U}_n(v)$ and $\mathcal{V}_n$ are such that for constants d_1 and d_2 it holds,

   $$P(V \in \mathcal{V}_n) \geq d_1, \quad P(U_{(v)} \in \mathcal{U}_n(v)) \geq d_2, \quad \forall v \in \mathcal{V}_n. \tag{2.12}$$

   Then the right side of (2.8) has a linear lower bound as desired:

   $$\psi_n \cdot \sum_{v \in \mathcal{V}_n} P(U_{(v)} \in \mathcal{U}_n(v))P(V = v) \geq (d_1\, d_2\, d\,)\, n.$$

3. The most crucial assumption of Theorem 2.1 is assumption **1)**. It states that the function $u \mapsto l(u, v)$ increases at least by certain amount δ on the set where U and V take their typical values. The core of the approach is to find U and V such that (2.6) holds. Later in concrete settings, we shall see how (2.6) is achieved in practice.

4. The condition **2)** barely states the existence of an uniform lower bound for the conditional variance (i.e. independent of v). Some trivial bounds clearly exist, but as explained above, ψ_n has to grow linearly in order to get a linear lower bound for $\mathrm{Var}[L_n]$.
5. The condition **3)** is of technical nature. In particular, it holds if $\mathcal{U}_n(v)$ is a lattice of span k_o, i.e. for integers m and u_o

$$\mathcal{U}_n(v) = \{u_o + k_o, u_o + 2k_o, \ldots, u_o + mk_o\}. \tag{2.13}$$

   As we shall see, this is a typical situation in practice.
6. The proof is based on the decomposition (2.9). In all previous papers, the lower bound of $\mathrm{Var}[L(Z)]$ was achieved by bounding below the (expectation) of conditional variance $\mathrm{Var}[L(Z)|U,V]$ ([8, 6, 20, 24, 37]). This approach often involves martingale's arguments (via Höffding-Azuma inequality), non-trivial combinatorial estimates and a generalization of Proposition 2.1. In this paper, however, we bound the variance of conditional expectation $\mathrm{Var}\big(E[L(Z)|U,V]\big)$. Although the main idea remains the same, the proof is now much shorter and less technical, relying solely on Proposition 2.1.

Corollary 2.3 *Let, for any $v \in \mathcal{V}_n$, the set $\mathcal{U}_n(v)$ be defined as $\mathcal{U}_n(v) = \mathcal{U}_n \cap \mathcal{S}_n(v)$, where $\mathcal{U}_n \subset \mathbb{R}$ is a subset independent of v. If the assumptions of Theorem 2.1 are satisfied, then*

$$\mathrm{Var}[L(Z)] \geq \left(\frac{\delta}{k_o}\right)^2 \cdot \psi_n \cdot P(U \in \mathcal{U}_n, V \in \mathcal{V}_n). \tag{2.14}$$

Proof. The (2.14) follows from (2.8):

$$\begin{aligned}\sum_{v\in\mathcal{V}_n} P\big(U_{(v)} \in \mathcal{U}_n(v)\big)P(V=v) &= \sum_{v\in\mathcal{V}_n} P\big(U_{(v)} \in \mathcal{U}_n\big)P(V=v) = \sum_{v\in\mathcal{V}_n} P\big(U \in \mathcal{U}_n|V=v\big)P(V=v)\\ &= \sum_{v\in\mathcal{V}_n} P\big(U \in \mathcal{U}_n, V=v\big) = P(U \in \mathcal{U}_n, V \in \mathcal{V}_n).\end{aligned}$$

■

2.3 Random transformation and the condition (2.6)

In order to simplify the notation, we consider the case where $\mathcal{U}_n$ is an integer interval and that, for any $v \in \mathcal{V}_n$, the fiber $\mathcal{S}_n(v)$ is a lattice with span $k_o \geq 1$. Thus, for every $v \in \mathcal{V}_n$ there exists an integer m (depending on n and v) so that $\mathcal{S}_n(v) \cap \mathcal{U}_n$ is as in (2.13). As explained in Remark 5, in this case the condition **3)** of Theorem 2.1 is fulfilled.

For any $(u,v) \in \mathcal{S}_n$, let $P_{(u,v)}$ denote the law of of Z given $U=u$ and $V=v$. Thus

$$P_{(u,v)}(z) = P(Z=z|U=u, V=v).$$

Recall from the introduction that the core of the whole two-step approach is the existence of a random transformation $\mathcal{R} : \mathcal{X}_n \times \mathcal{X}_n \to \mathcal{X}_n \times \mathcal{X}_n$ independent of Z that satisfies (1.5). In order to make this approach to work, the transformation should be associated with the U and V in the following way: for a typical $z \in \mathcal{X}_n \times \mathcal{X}_n$, the transformation increases $\mathfrak{u}(z)$ by k_o unit and leaves $\mathfrak{v}(z)$ unchanged. Typically there are many such (random or non-random) mappings, but to ensure (2.6), the transformation should be chosen so that some additional assumptions are fulfilled. Recall that $\tilde{Z}$ is obtained from Z by applying a random modification to Z and the additional randomness is independent of Z. As mentioned above, the transformation increases $U = \mathfrak{u}(Z)$ by k_o and leaves $V = \mathfrak{v}(Z)$ unchanged, thus (at least for the typical values of Z), it holds

$$\mathfrak{u}(\tilde{Z}) = \mathfrak{u}(Z) + k_o, \quad \mathfrak{v}(\tilde{Z}) = \mathfrak{v}(Z). \tag{2.15}$$

In addition, we need the following assumptions to be true:

A1 There exist universal (not depending on n) constants $\alpha > 0$ and $\epsilon_o > 0$ such that

$$P\big(E[L(\tilde{Z}) - L(Z)|Z] \geq \epsilon_o\big) \geq 1 - \exp[-n^\alpha].$$

A2 There exists universal constant $A < \infty$ so that $L(\tilde{Z}) - L(Z) \geq -A$.

A3 For any (u, v) such that $v \in \mathcal{V}_n$ and $u \in \mathcal{U}_n(v)$, the following implication holds:

$$\text{If } Z \sim P_{(u,v)}, \text{ then } \tilde{Z} \sim P_{(u+k_o,v)}. \tag{2.16}$$

Remarks:

1. The assumption **A1** is the condition (1.5) explained already in Introduction.
2. The assumption **A2** states that by applying the random transformation, the maximum decrease of the score is at most A. This assumption usually holds for trivial reasons.
3. Note that (2.16) implies (2.15). However, the condition (2.16) is more restrictive and (except some trivial cases) to achieve it, the transformation R has to be random.
4. If $\mathcal{U}_n(v) = \mathcal{U}_n \cap \mathcal{S}_n(v)$, then $v \in \mathcal{V}_n$ and $u \in \mathcal{U}_n(v)$ holds if and only if $(u, v) \in \mathcal{S}_n \cap (\mathcal{U}_n \times \mathcal{V}_n)$.

Theorem 2.2 *Assume the existence of a random transformation so that for every n,* **A1**, **A2** *and* **A3** *hold. Suppose that there exists a constant $a > 0$ so that for any (u, v) such that $v \in \mathcal{V}_n$ and $u \in \mathcal{U}_n(v)$, it holds*

$$P(U = u, V = v) \geq \frac{1}{an}. \tag{2.17}$$

Then there exists a $n_5 < \infty$ so that for every $n > n_5$ the assumption **1)** *of Theorem 2.1 is fulfilled with $\delta = \frac{\epsilon_o}{2}$.*

Proof. Let $v \in \mathcal{V}_n$ and $u \in \mathcal{U}_n(v)$. Let $Z_{(u,v)}$ be a random vector having the distribution $P_{(u,v)}$. By **A3**, thus,

$$l(u + k_o, v) = E[L(\tilde{Z}_{(u,v)})].$$

Hence

$$\begin{aligned} l(u + k_o, v) - l(u, v) &= E[L(\tilde{Z}_{(u,v)})] - E[L(Z_{(u,v)})] = E[L(\tilde{Z}_{(u,v)}) - L(Z_{(u,v)})] \\ &= E\big(E[L(\tilde{Z}_{(u,v)}) - L(Z_{(u,v)})\big|Z_{(u,v)}]\big). \end{aligned}$$

Let $B_n \subset \mathcal{X}_n \times \mathcal{X}_n$ be the set of outcomes of Z such that

$$\big\{E[L(\tilde{Z}) - L(Z)|Z] \geq \epsilon_o\big\} = \{Z \in B_n\}.$$

By assumption **A2**, for any pair of sequences z, the worst decrease of the score, when applying the block-transformation is $-A$. Hence

$$E\big(E[L(\tilde{Z}_{(u,v)}) - L(Z_{(u,v)})\big|Z_{(u,v)}]\big) \geq \epsilon P\big(Z_{(u,v)} \in B_n(\epsilon)\big) - AP(Z_{(u,v)} \notin B_n(\epsilon)).$$

By **A1**, $P\big(Z \notin B_n\big) \leq \exp[-n^\alpha]$. Therefore

$$P\big(Z_{(u,v)} \notin B_n\big) = P\big(Z \notin B_n|U = u, V = v\big) \leq \frac{P\big(Z \notin B_n\big)}{P(U = u, V = v)} \leq an \exp[-n^\alpha]$$

where the last inequality follows from (2.17). Take now n_5 so big that for any $n > n_5$, we have

$$\epsilon_o\big(1 - an \exp[-n^\alpha]\big) - Aan \exp[-n^\alpha] > \frac{\epsilon_o}{2}.$$

Hence, for any $n > n_5$ and for any (u, v) such that $v \in \mathcal{V}_n$ and $u \in \mathcal{U}_n(v)$, we have

$$l(u + k_o, v) - l(u, v) \geq \epsilon_o\big(1 - an \exp[-n^\alpha]\big) - Aan \exp[-n^\alpha] \geq \frac{\epsilon_o}{2}. \tag{2.18}$$

■

2.4 Covered previous results

Before turning into new results presented in the subsequent sections, let us briefly mention how the random transformation as well as the associated random variables were defined in already obtained results:

- In [8], the random variable U is the number of matching replica points while V is a constant. Roughly speaking, a letter X_i is a replica point if it has a neighborhood that matches exactly with the periodic sequence (i.e. it has the same periodic pattern). The replica point itself can or cannot match, and it is shown that the number of matching replica points has variance proportional to n. In [8] the random transformation is not explicitly defined, but one can take it as uniformly choosing a replica point with prescribed color and change its value.

- In [6] the random sequence X is built up on the alphabet $\{0, 1, a\}$. The random variable U is the number of a's in X, while V is a constant. The random transformation is hidden in the so called *drop-scheme of random bits*, used to construct the sequence X^{01} which is the subsequence of X only having 0's and 1's. Roughly speaking, the drop-scheme of random bits consists on, starting from a binary random sequence of length two, to flip a coin and to add the resulting symbol into the previous sequence at an uniformly chosen location, so increasing the length, until reaching a length $n - U$.
- In [20] the scoring function is such that $S(1, 0) = S(0, 1) = 0, S(0, 0) = 1$ and $S(1, 1) \in \mathbb{R}$. The random transformation consists, in X, to uniformly choose a block of length five, to take one of its symbols out and to add it to a uniformly chosen block of length one. The random variable U is the number of blocks of length two and of length four, and V is the number of blocks of the other lengths.
- In [24], both sequences typically consist of many zeros and few ones. The random transformation, uniformly at random picks an arbitrary one in Z and turns it into a zero. The variable U is the number of ones in Z, V is a constant. Hence, this case is essentially the same as considered in the Section 3 with $|\mathbb{A}| = 2$ and Theorem 3.2 nicely generalizes Theorem 2.1 in [24].
- In [37], the random transformation and the random variables (U, V) are defined to be as in Section 4.

3 Optimal score of random i.i.d. sequences

Our first application deals with the general scoring scheme as introduced in Section 1.1. Thus let $\mathbb{A}$ be a finite alphabet and X, Y be independent i.i.d. sequences so that any letter has positive probability, i.e.

$$P(c) := P(X_1 = c) > 0, \quad \forall c \in \mathbb{A}.$$

Clearly now $\mathcal{X}_n = \mathbb{A}^n$. Let $S : \mathbb{A} \times \mathbb{A} \to \mathbb{R}^+$ be a scoring function. Let $A < \infty$ be the maximal value of the scoring function, i.e. $\max_{a,b} S(a, b) \leq A$. We naturally assume that the gap price does not exceed A, i.e. $\delta \leq A$. Now, it is easy but important to observe that changing one letter in the sequence X, say X_1, decreases the score at most by A units. Indeed, if X_1 was not included any optimal alignment, then changing it does not decrease the score. If an optimal alignment includes X_1, then after the change, the previous alignment (which now need not to be optimal any more) scores at most A units less than before the change. And the new optimal alignment cannot score less.

The random transformation. In this setup, the random transformation is the following. Recall that Z stands for the pair of sequences (X, Y). We choose two specific letters a and b from the alphabet $\mathbb{A}$. Given the pair Z such that at least one of the sequences contain at least one a, we choose a letter a from Z with uniform distribution and change

it into the letter b. Hence $\tilde{Z}$ and Z differ from one letter only and as just explained, the maximum decrease of score is at most A, i.e. $L(\tilde{Z}) - L(Z) \geq -A$ i.e. the condition **A2** is satisfied. Choosing such a transformation is motivated by the following result (c.f. Theorem 5.1 and Theorem 5.2 in [26]):

Theorem 3.1 *Suppose there exist letters $a, b \in \mathbb{A}$ such that*

$$\sum_{c \in \mathbb{A}} P(c)\big(S(b,c) - S(a,c)\big) > 0. \tag{3.1}$$

Then, there exist constants $\delta_0 > -\infty$, $\epsilon_o > 0$, $\alpha > 0$ and $n_0 < \infty$ such that

$$P\big(E[L(\tilde{Z}) - L(Z)|Z] \geq \epsilon_o\big) \geq 1 - e^{-\alpha n} \tag{3.2}$$

given $\delta < \delta_0$ and $n \geq n_0$.

Remarks:

1. For the two-letter alphabet $\mathbb{A} = \{a, b\}$, condition (3.1) says

   $$\big(S(b,a) - S(a,a)\big)P(a) + \big(S(b,b) - S(b,a)\big)P(b) > 0.$$

   When $S(b,b) = S(a,a) \neq S(a,b) = S(b,a)$, then (3.1) is satisfied if and only if $P(a) \neq P(b)$. For, example when $S(b,b) = S(a,a) > S(b,a) = S(b,a)$, then (3.1) holds if $P(a) < P(b)$.

2. The condition $\delta < \delta_0$ means that the gap penalty $-\delta$ has to be sufficiently large. Intuitively, the larger the gap penalty (smaller the gap price), the less gaps in optimal alignment so that the optimal alignment is closer to the pairwise comparison (Hamming distance). Some methods for determining a sufficient δ_0, as well as some examples, are discussed in [26]. We believe that the assumption on δ can be relaxed so that Theorem 3.1 holds under more general assumptions.

In this section, we shall assume that there exists letters $a, b \in \mathbb{A}$ so that the random transformation satisfies **A1** (equivalently, (3.2)). We shall show that all other assumptions of Theorems 2.1 and 2.2 are fulfilled. We start with the general case $|\mathbb{A}| > 2$, the case $|\mathbb{A}| = 2$ will be discussed in the end of the present section.

3.1 The case $|\mathbb{A}| > 2$

Let $\mathbb{A} = \{a, b, c_1, \ldots, c_l\}$, where $l \geq 1$. The letters a and b are the ones used in the random transformation. Let

$$q_j := \frac{P(c_j)}{1 - P(a) - P(b)}, \quad j = 1, \ldots, l.$$

With N_a and N_b being the random number of a's and b's in $X_1, \ldots, X_n, Y_1, \ldots, Y_n$, we define the random variables

$$U := N_b, \quad V := N_a + N_b.$$

For any $z \in \mathcal{X}_n \times \mathcal{X}_n$, thus $\mathfrak{u}(z)$ and $\mathfrak{v}(z) - \mathfrak{u}(z)$ stand for the number of b's and the number of a's in both strings, respectively. The random transformation applied on z changes an randomly chosen a into a letter b, hence the transformation increases $\mathfrak{u}(z)$ by one, whilst $\mathfrak{v}(z)$ remains unchanged.
Clearly the possible values for U and V are $\{0, \ldots, 2n\}$ and the only restriction to (U, V) is that $U \leq V$. Hence, in this case $\mathcal{S}_n^U = \mathcal{S}_n^V = \{0, \ldots, 2n\}$,

$$\mathcal{S}_n := \{(u, v) \in \{0, \ldots, 2n\} \times \{0, \ldots, 2n\} : u \leq v\}$$

and for any v,

$$\mathcal{S}_n(v) = \{0, \ldots, v\}.$$

For any $z \in \mathcal{X}_n \times \mathcal{X}_n$,

$$\begin{aligned} P(Z = z | U = u, V = v) &= P(Z = z | N_a = v - u, N_b = u) \\ &= \begin{cases} \prod_{j=1}^{l} q_j^{m_j(x)} \binom{2n}{u \; v-u \; 2n-v}^{-1}, & \text{if } \mathfrak{u}(z) = u, \mathfrak{v}(z) = v \\ 0, & \text{else} \end{cases}, \end{aligned} \tag{3.3}$$

where $m_j(z)$ is the number of c_j-colored letters in z.

The sets $\mathcal{U}_n(v)$ and $\mathcal{V}_n$. Note that $U \sim B(2n, P(b))$ and $V \sim B(2n, P(a) + P(b))$. Also note that for any $v > 0$,

$$U_{(v)} \sim B(v, p_b),$$

where $p_b = \frac{P(b)}{P(a)+P(b)}$. Let $p := P(a) + P(b)$ and let

$$\begin{aligned} \mathcal{V}_n &:= \left[2np - \sqrt{2n}, 2np + \sqrt{2n}\right] \cap \mathcal{S}_n^V, \\ \mathcal{U}_n(v) &:= \left[vp_b - \sqrt{v}, vp_b + \sqrt{v}\right] \cap \mathcal{S}_n(v). \end{aligned}$$

Now it is clear that the condition **3)** of Theorem 2.1 is fulfilled with $k_o = 1$.

With the help of Chebyshev's inequality, it is straightforward to see that for any n,

$$P(V \in \mathcal{V}_n) \geq 1 - p(1-p), \quad P(U_{(v)} \in \mathcal{U}_n(v)) \geq 1 - p_b(1-p_b). \tag{3.4}$$

Clearly, there exists $v_o(p_b)$ so that $vp_b + \sqrt{v} < v$, whenever $v > v_o$. Thus, there exists n_1 so that for every $n > n_1$ and $v \in \mathcal{V}_n$, it holds that $v > v_o$. In particular, for every $n > n_1$ and for every pair (u, v) such that $v \in \mathcal{V}_n$, $u \in \mathcal{U}_n(v)$, it holds $v > u$.

Lemma 3.1 *There exist an universal constant $a > 0$ and $n_2 > n_1$ such that for any $n > n_2$, for any $v \in \mathcal{V}_n$ and $u \in \mathcal{U}_n(v)$, it holds*

$$P(U = u, V = v) \geq \frac{1}{an}. \tag{3.5}$$

Proof. The proof is based on Lemma 2.3. Since $U_{(v)} \sim B(v, p_b)$, by (2.3) there exists v_o and a positive constant b_1 so that for any $v > v_o$,

$$P(U_{(v)} = u) \geq \frac{1}{b_1\sqrt{v}}, \quad \forall u \in \mathcal{U}_n(v). \tag{3.6}$$

Secondly, since $V \sim B(2n, p)$, there exists $n_{2,1}$ so that for every $n > n_{2,1}$,

$$P(V = v) \geq \frac{1}{b_2\sqrt{2n}}, \quad \forall v \in \mathcal{V}_n. \tag{3.7}$$

Take now $n_2 > n_{2,1}$ so big that for any $n > n_2$ it holds $2np - \sqrt{2n} > v_o$. Then, for every $v \in \mathcal{V}_n$, (3.6) and (3.7) both hold. Thus, for any $n > n_2$, $v \in \mathcal{V}_n$ and $u \in \mathcal{U}_n(v)$, we have

$$P(U = u, V = v) = P(U_{(v)} = u)P(V = v) \geq \frac{1}{b_1\sqrt{v}b_2\sqrt{2n}} \geq \frac{1}{(b_1b_2)\sqrt{2n(2np + \sqrt{2n})}} \geq \frac{1}{an},$$

where the constant a can be taken as $2b_1b_2\sqrt{p+1}$.

■

Lemma 3.2 *There exists a finite n_3 and a constant $d > 0$ such that $n_3 > n_2$ and for every $n > n_3$ and $v \in \mathcal{V}_n$, it holds*

$$\mathrm{Var}[U_{(v)} | U_{(v)} \in \mathcal{U}_n(v)] \geq dn. \tag{3.8}$$

Proof. From Lemma 2.3, we know that there exists c_1 and v_o, so that

$$\mathrm{Var}[U_{(v)} | U_{(v)} \in \mathcal{U}_n(v)] \geq c_1 v, \tag{3.9}$$

provided $v > v_o$. Let $n_{3,1}$ be such that for every $n > n_{3,1}$ $2np - \sqrt{2n} > v_o$. Then, for any $n > n_{3,1}$ an any $v \in \mathcal{V}_n$, we have that $v > v_o$ so that (3.9) holds and

$$\mathrm{Var}[U_{(v)} | U_{(v)} \in \mathcal{U}_n(v)] \geq c_1(2pn - \sqrt{2n}). \tag{3.10}$$

Finally take $n_3 > n_{3,1}$ so big that for a constant $d > 0$, $c_1(2pn - \sqrt{2n}) \geq dn$, provided $n > n_3$.

■

Finally we prove **A3** for that particular model.

Lemma 3.3 *Let $(u, v) \in \mathcal{S}_n$ be such that $v > u$. Let $Z \sim P_{(u,v)}$ Then $\tilde{Z} \sim P_{(u+1,v)}$*

Proof. For any $z \in \mathcal{X}_n \times \mathcal{X}_n$, let the set $\mathcal{A}(z)$ consists of possible outcomes after applying the random transformation to z. Since the transformation changes an a into b, the number of different outcomes equals to the number of a's in z, thus $|\mathcal{A}(z)| = \mathfrak{v}(z) - \mathfrak{u}(z)$. Since the transformation picks the letters uniformly, we obtain that for any $\tilde{z} \in \mathcal{A}(z)$,

$$P(\tilde{Z} = \tilde{z} | Z = z) = \frac{1}{\mathfrak{v}(z) - \mathfrak{u}(z)}. \tag{3.11}$$

Let the set $\mathcal{B}(\tilde{z})$ consist of all these pairs of strings that could result $\tilde{z}$ after the transformation: $\mathcal{B}(\tilde{x}) := \{x \in \mathcal{X} : \tilde{x} \in \mathcal{A}(x)\}$. Since before transformation one of b's in $\tilde{z}$ was a, clearly $|\mathcal{B}(\tilde{z})| = \mathfrak{u}(\tilde{z})$. Recall that U and V are the functions of Z. Let $Z \sim P_{(u,v)}$. We aim to find

$$P(\tilde{Z} = \tilde{z}) = P(\tilde{Z} = \tilde{z} | U = u, V = v) = \sum_{z \in \mathcal{B}(\tilde{z})} P(\tilde{Z} = \tilde{z} | Z = z) P(Z = z | U = u, V = v). \tag{3.12}$$

Let

$$S(u, v) := \{z : \mathfrak{u}(z) = u, \mathfrak{v}(z) = v\}.$$

Clearly the right hand side of (3.12) is zero, if

$$S(u, v) \cap \mathcal{B}(\tilde{z}) = \emptyset.$$

This simply means that the string $\tilde{z}$ does not satisfy at least one of the following equalities:

$$\mathfrak{u}(\tilde{z}) = u + 1, \quad \mathfrak{v}(\tilde{z}) = v.$$

Let us now assume that $\tilde{z}$ satisfies both equalities above. In particular, $|\mathcal{B}(\tilde{z})| = u + 1$ and any element in $\mathcal{B}(\tilde{z})$ is such that the number of b's is u and the number of a's is $v - u$ and the number of all c_j equal to that of $\tilde{z}$. i.e. $m_j(z) = m_j(\tilde{z})$ $\forall \tilde{z} \in \mathcal{B}(\tilde{z})$. Clearly $\mathcal{B}(\tilde{z}) \subset S(u, v)$. By (3.3), thus

$$\begin{aligned}
P(\tilde{Z} = \tilde{z} | U = u, V = v) &= \sum_{z \in \mathcal{B}(\tilde{z})} P(\tilde{Z} = \tilde{z} | Z = z) P(Z = z | U = u, V = v) \\
&= \frac{1}{v - u} |\mathcal{B}(\tilde{z})| \binom{2n}{v - u \; u \; 2n - v}^{-1} \prod_{j=1}^{l} q_j^{m_j(\tilde{z})} \\
&= \frac{u + 1}{v - u} \binom{2n}{v - u \; u \; 2n - v}^{-1} \prod_{j=1}^{l} q_j^{m_j(\tilde{z})} \\
&= \frac{u!(u + 1)(v - u)!(2n - v)!}{(v - u)2n!} \prod_{j=1}^{l} q_j^{m_j(\tilde{z})} \\
&= \frac{(u + 1)!(v - u - 1)!(2n - v)!}{2n!} \prod_{j=1}^{l} q_j^{m_j(\tilde{z})} \\
&= \prod_{j=1}^{l} q_j^{m_j(\tilde{z})} \binom{2n}{v - u - 1 \; u + 1 \; 2n - v}^{-1}.
\end{aligned}$$

By (3.3), $\tilde{Z} \sim P_{(u+1,v)}$.

■

Theorem 3.2 *Assume that the random transformation satisfies* **A1**. *Then there exists an universal constant $b > 0$ and $n_6 < \infty$ so that for every $n > n_6$, it holds*

$$\mathrm{Var}[L(Z)] \geq b \cdot n. \tag{3.13}$$

Proof. Let us first check the assumptions of Theorem 2.2. **A1** holds by the assumption. As explained above, the random transformation is such that **A2** holds. Let now n_2 be as in Lemma 3.1 and n_3 as in Lemma 3.2. Recall that $n_1 < n_2 < n_3 < \infty$. Hence, for any $n > n_3$, (3.5) and (3.8) hold; moreover, from (3.4), it follows that with $d_1 = 1 - p(1-p)$ and $d_2 = 1 - p_b(1-p_b)$, the inequalities (2.12) hold and for any pair (u, v) where $v \in \mathcal{V}_n$ and $u \in \mathcal{U}_n(v)$, we have that $v > u$. The latter ensures that the random transformation is possible and Lemma 3.3 now establishes **A3**. Therefore, for every $n > n_3$, the assumptions of Theorem 2.2 are fulfilled and so there exists $n_5 > n_3$ so that for for every $n > n_5$, the assumptions of Theorem 2.1 hold with $\delta = \frac{\epsilon_0}{2}$.
We now apply Theorem 2.1. As just showed, the assumption **1**) holds for any $n > n_5$; as explained in Subsection 3.1, the assumption **3)** holds with $k_o = 1$. By (3.8), $\psi_n = dn$. By Theorem 2.1, thus, the lower bound (4.20) exists with

$$b = \frac{\epsilon_o^2 d_1 d_2 d}{4}.$$

■

3.2 The case $A = \{a, b\}$

This case is easier. The only random variable is U, formally we can take $V \equiv 2n$. Then (3.3) is

$$P(Z = z | U = u) = P(Z = z | U = u, V = 2n) = \binom{2n}{u}^{-1} I_{\{\mathfrak{u}(z)=u\}}.$$

Now take $\mathcal{V}_n = \{2n\}$ and

$$\mathcal{U}_n = \mathcal{U}_n(2n) = [2np_b - \sqrt{2n}, 2np_b - \sqrt{2n}] \cap S_n^U.$$

Then everything holds true: there clearly exists n_1 so that $u < v = 2n$, whenever $n > n_1$ and $u \in \mathcal{U}_n$; the bound (3.5) holds with $a = b_1$ (and, in fact $n^{-\frac{1}{2}}$ instead of n^{-1}); the proof of Lemma 3.2 is simply (3.9) and the proof of Lemma 3.3 holds true with $\mathfrak{v}(z) = 2n$. Thus Theorem 3.2 holds.

4 The length of the LCS of random i.i.d. block sequences

In this section we are interested in the fluctuations of $L(Z)$ for the score function as in (1.2), where $Z = (X, Y)$ are binary sequences having a certain random block structure. This random block structure was first considered in [37]. In the present article, we consider a random block structure which is a generalization of the model in [37]. We are able to show that the length of the longest common subsequence of two sequences having this random block structure grows linearly by following the general two-step approach. Therefore, we confirm in this setting Waterman's conjecture.

Ofte in this section x, y stand for binary strings of length $n > 0$. We start by getting a bit more familiar with the LCS of x and y. First, note that the LCS of x and y and the alignment generating it are not necessarily unique, but its length is unique.

Example 4.1 *Let* $x = 100101100001101$ *and* $y = 111000010101110$. *The length of the LCS of* x *and* y *is* $L_{15}(x, y) = 10$. *A candidate for the LCS of* x *and* y *is the string* 1000100111. *This LCS could have came (but not exclusively) from any of the following alignments:*

x		1	−	−	0	0	1	0	1	−	1	0	0	0	−	0	1	1	−	0	1	−
y		1	1	1	0	0	−	0	−	0	1	−	0	−	1	0	1	1	1	−	1	0
LCS		**1**			**0**	**0**		**0**			**1**		**0**			**0**	**1**	**1**			**1**	

x		1	−	−	0	0	1	0	1	−	1	0	0	0	−	0	1	−	1	0	1	−
y		1	1	1	0	0	−	0	−	0	1	−	−	0	1	0	1	1	1	−	1	0
LCS		**1**			**0**	**0**		**0**			**1**			**0**		**0**	**1**		**1**		**1**	

Another candidate for an LCS of x *and* y *is the string* 1000000110.

We now introduce the random block model:

4.1 The 3-multinomial block model

We say that a block of zeros of length $m \in \mathbb{Z}_+$ in x is a run of 0's of maximal length between two ones, except for the block of zeros at the beginning of x, which only has a 1 inmediatly to its left. We consider the analog convention for a block of ones of length m in x, as well as for any binary sequence. Let $l \geq 2$ and $q_1, q_2, q_3 \in (0, 1)$ be parameters such that $q_1 + q_2 + q_3 = 1$. Let $(W_k)_k$ and $(W'_k)_k$ be two i.i.d. sequences of random variables taking values on $\{l-1, l, l+1\}$, independent of each other, with distribution

$$P(W_k = l_i) = P(W'_k = l_i) = q_i\,, \qquad \forall\, k \geq 1,\ i \in \{1, 2, 3\}$$

where $l_1 := l - 1, l_2 := l$ and $l_3 := l + 1$. Let $(w_k)_k$ be a realization of $(W_k)_k$. Let us construct $x^\infty = x_1 x_2 x_3 \cdots$ an infinite binary sequence depending on $(w_k)_k$ as follows:

We choose $\vartheta \in \{0,1\}$ with probability $1/2$, independently from everything else. Then, we built up the first block in x^∞ as a block of ϑ's with lenght w_1, the second block in x^∞ as a block of $(1-\vartheta)$'s with length w_2, the third block in x^∞ as a block of ϑ's with length w_3, and so on. We built up, in a completely analog way, the sequence y^∞ based on a realization $(w'_k)_k$ of $(W'_k)_k$. After this, for a given $n > 0$, let us define $\hat{x}((w_k)_k) := x^\infty[1,n] := x_1 \cdots x_n$ and $\hat{y}((w'_k)_k) := y^\infty[1,n] := y_1 \cdots y_n$, namely the first n-bytes of the infinite sequence x^∞, respectively y^∞. Note that the last run of the same symbol in $\hat{x}((w_k)_k)$ (resp. in $\hat{y}((w'_k)_k)$) is not a block according to our definition, or though its length r is such that $r \in \{1,\ldots,l+1\}$. Naturally, $\hat{x}((w_k)_k)$ (or equivalently $\hat{y}((w'_k)_k)$) induces the set $\mathcal{X}_n \subset \{0,1\}^n$ of binary sequences having blocks only with length either $l-1, l$ or $l+1$ and a last run of the same symbol with length r such that $r \in \{1,\ldots,l+1\}$. Let us denote by $X := \hat{x}((W_k)_k) := X_1 \cdots X_n$ (resp. by $Y := \hat{y}((W'_k)_k) := Y_1 \cdots Y_n$) the associated random binary sequence of length n whose realization is an element of $\mathcal{X}_n$. The process of allocating the blocks can be seen as independently drawing balls of 3 different colours from an urn, where a ball of colour i has probability q_i to be picked up, $i = 1,2,3$. That is why we call this the 3-multinomial block model. For $k \in \{l-1, l, l+1\}$ and $x \in \mathcal{X}_n$, let $b_k(x)$ be the number of blocks of length k in x, and denote $B_k := b_k(X)$ the associated random variable.

Example 4.2 *Take $l = 3$ and let $W_1 = 2, W_2 = 3, W_3 = 2, W_4 = 4, W_5 = 3, \ldots$ Suppose that we get $\vartheta = 0$, so then $x^\infty = 00111001111000\ldots$. For $n = 13$, we get the sequence $x = 0011100111100$ such that $b_2(x) = 2, b_3(x) = 1$ and $b_4(x) = 1$, with a rest at the end of length $r = 2$.*

Let us define the following three new random variables:

$$T := B_l + B_{l-1} + B_{l+1} \tag{4.1}$$

$$U := B_l - B_{l-1} - B_{l+1} \tag{4.2}$$

$$R := n - (\, l\, B_l + (l+1)\, B_{l+1} + (l-1)\, B_{l-1}\,). \tag{4.3}$$

Given $(b_{l-1}(x), b_l(x), b_{l+1}(x))$, let us denote by $(t(x), u(x), r(x))$ the solution of the linear system

$$\begin{aligned} t(x) &= b_l(x) + b_{l-1}(x) + b_{l+1}(x) \\ u(x) &= b_l(x) - b_{l-1}(x) - b_{l+1}(x) \\ r(x) &= n - (\, l\, b_l(x) + (l+1)\, b_{l+1}(x) + (l-1)\, b_{l-1}(x)\,). \end{aligned}$$

The other way around, given any realization $(t,u,r) \in \mathbb{Z}^3$ of (T,U,R), let us denote by

$$(b_{l-1}(t,u,r), b_l(t,u,r), b_{l+1}(t,u,r))$$

the solution of the linear system:

$$\begin{pmatrix} b_{l-1}(t,u,r) \\ b_l(t,u,r) \\ b_{l+1}(t,u,r) \end{pmatrix} = \begin{pmatrix} (2l+1)/4 & -1/4 \\ 1/2 & 1/2 \\ -(2l-1)/4 & -1/4 \end{pmatrix} \begin{pmatrix} t \\ u \end{pmatrix} + \begin{pmatrix} -(n-r)/2 \\ 0 \\ (n-r)/2 \end{pmatrix}. \tag{4.4}$$

This means that we have a one-to-one correspondence between the random variables (B_{l-1}, B_l, B_{l+1}) and (T,U,R), which will be often used in what follows.

The 3-multinomial distribution. We can compute the distribution of X by taking into account its block structure. In order to do so, let us define the function

$$p(r) := P(W \geq r), \qquad \text{for } r \in \{1, \ldots, l+1\}.$$

Clearly, $p(r) = 1$ when $r \in \{1, \ldots, l-1\}$, but $p(l) = q_2 + q_3$ and $p(l+1) = q_3$. Now, we see that for any $x \in \mathcal{X}_n$ it holds

$$P(X = x) = \frac{1}{2} q_1^{b_1(x)} q_2^{b_2(x)} q_3^{b_3(x)} p(r(x)), \tag{4.5}$$

where the factor $\frac{1}{2}$ is needed because to every fixed block-sequence corresponds two sequences in $\mathcal{X}_n$, both having the same probability (it is the choosing of the colour of the first block with probability 1/2). Moreover, e.g. by the urn analogy, we find that the joint distribution of (B_{l-1}, B_l, B_{l+1}) can be computed as following: given $b_1, b_2, b_3 \in \mathbb{N}$ it holds

$$P(B_{l-1} = b_1, B_l = b_2, B_{l+1} = b_3) = \sum_{x \in S(b_1,b_2,b_3)} P(X = x) = \binom{b_1 + b_2 + b_3}{b_1 \;\; b_2 \;\; b_3} q_1^{b_1} q_2^{b_2} q_3^{b_3} p(r), \tag{4.6}$$

where

$$S(b_1, b_2, b_3) := \{x \in \mathcal{X}_n : b_{l-1}(x) = b_1, b_l(x) = b_2, b_{l+1}(x) = b_3\}$$

and $r = n - (l-1)b_1 - lb_2 - (l+1)b_3$. Note that the factor $\frac{1}{2}$ disappears. So, combining (4.5) and (4.6) we naturally get that for $x \in \mathcal{X}$ it holds

$$P(X = x | B_{l-1} = b_1, B_l = b_2, B_{l+1} = b_3) = \frac{1}{2} \binom{b_1 + b_2 + b_3}{b_1 \;\; b_2 \;\; b_3}^{-1} 1_{S(b_1,b_2,b_3)}(x). \tag{4.7}$$

Note also that, from (4.4), we can even compute the joint distribution of (T, U, R) as follows:

$$P(U = u, T = t, R = r) = \binom{t}{b_{l-1}(t,u,r) \; b_l(t,u,r) \; b_{l+1}(t,u,r)} q_1^{b_{l-1}(t,u,r)} q_2^{b_l(t,u,r)} q_3^{b_{l+1}(t,u,r)} p(r). \tag{4.8}$$

4.2 Fluctuations of the length of the LCS in the 3-multinomial block model

Let $Z = (X, Y)$ be a vector of binary sequences, where each component has the previously introduced random block structure. Let us identify U defined in (4.2) with the random variable $\mathfrak{u}(Z)$ as well as the vector (T, R) with the random variable $\mathfrak{v}(Z)$. Therefore, in what follows $(U, V) = (T, U, R)$ and its support $\mathcal{S}_n$ consists of triples (t, u, r) so that $P(U = u, T = t, R = r) > 0$. We would like to use Theorem 2.1, so we must look for sets

$\mathcal{U}_n$ and $\mathcal{V}_n$ such that the conditions **1), 2)** and **3)** are satisfied. For any $c > 0$, define

$$\begin{aligned}
\mathcal{U}_n^c &:= \left[\frac{n}{\mu}(q_2 - q_1 - q_3) - c\sqrt{n}, \frac{n}{\mu}(q_2 - q_1 - q_3) + c\sqrt{n}\right] \cap \mathbb{Z}, \\
\mathcal{T}_n^c &:= \left[\frac{n}{\mu} - c\sqrt{n}, \frac{n}{\mu} + c\sqrt{n}\right] \cap \mathbb{Z}, \\
\mathcal{V}_n^c &:= \big(\mathcal{T}_n^c \times \{1, \dots, l+1\}\big) \cap \mathcal{S}_n^V.
\end{aligned}$$

Note that the notation $\mathcal{U}_n^c$ (resp. $\mathcal{T}_n^c$ and $\mathcal{V}_n^c$) means that the set $\mathcal{U}_n$ explicitely depends on the constant $c > 0$, and has nothing to do with the notation for the complement of a set. Recall the right hand side of (2.14) and the 2. Remark after Theorem 2.1. A first observation is that, uniformly on n, $P(U \in \mathcal{U}_n^c, V \in \mathcal{V}_n^c)$ is bounded by below by a constant:

Lemma 4.1 *There exist universal constant $c > 0$ (not depending on n) and $n_0 < \infty$ such that for every $n > n_0$ it holds*

$$P\big(U \in \mathcal{U}_n^c, V \in \mathcal{V}_n^c\big) \geq 0.9.$$

The proof is a rather straightforward application of large deviation techniques, and therefore it is contained in the Appendix. We shall also need the following lemma (proven in the Appendix):

Lemma 4.2 *There exist an universal constant $\alpha > 0$ and $n_1 \in (n_0, \infty)$ such that for every $n > n_1$ and $(u, v) \in \mathcal{S}_n \cap (\mathcal{U}_n \times \mathcal{V}_n)$, it holds*

$$\left| b_{l_i}(u, v) - q_i \frac{n}{\mu} \right| \leq \alpha\sqrt{n} \quad \textit{and} \quad b_{l_i}(u, v) \geq 1, \quad \forall\, i = 1, 2, 3. \tag{4.9}$$

In what follows, we shall take $\mathcal{U}_n := \mathcal{U}_n^c$ and $\mathcal{V}_n = \mathcal{V}_n^c$, where $c > 0$ is as in Lemma 4.1 and we shall take $n > n_0$. Recall the definition $\mathcal{U}_n(v) := \mathcal{U}_n \cap \mathcal{S}_n(v)$, for $v \in \mathcal{V}_n$. We show now how the conditions of Theorem 2.1 are fulfilled:

Condition 3). Lets us start by showing that $u \in \mathcal{U}_n(v)$ implies $u + i \notin \mathcal{U}_n(v)$ for $i = 1, 2, 3$. Indeed given $u \in \mathcal{U}_n(v)$, it is enough to realize that $b_{l-1}(u + i, v)$ is not an integer if $i = 1, 2, 3$ but $b_{l-1}(u + 4, v)$ is an integer, where $b_{l-1}(u, v), b_l(u, v), b_{l+1}(u, v)$ are the integer solutions of the system (4.4). Next, from Lemma 4.2, it follows (among other things) that, if n is big enough, $u+4 \in \mathcal{S}_n(v)$ for every $v \in \mathcal{V}_n$ and $u \in \mathcal{U}_n(v)$, which finally implies that $\mathcal{U}_n(v)$ is of the form (2.13) with $k_o = 4$ so that the condition **3**) of Theorem 2.1 holds with $k_o = 4$. Let us explain all the last argument. For every $n > n_1$, $v \in \mathcal{V}_n$ and $u \in \mathcal{U}_n(v)$, we have that necessarily $u + 4 \in S_n(v)$, so $(u, v) \in \mathcal{S}_n$, therefore there exists at least one possible outcome $x \in \mathcal{X}_n$ so that $\mathfrak{u}(x) = u$ and $\mathfrak{v}(x) = v$. Moreover, from (4.9),

it follows that $b_{l_i}(u,v) \geq 1$ for every $i = 1,2,3$. Thus, in x there are at least one block from every size $\{l-1, l, l+1\}$. Deleting from x one bit from a block of the length $l+1$ and adding one bit to the block of length $l-1$, we turn both them to the blocks of the length l. This transformation does not change the number of blocks and the sum of the lengths of all blocks, so we have another possible outcome of X, say $\tilde{x}$ with $\mathfrak{u}(\tilde{x}) = u+4$ and $\mathfrak{v}(\tilde{x}) = v$. Hence $(u+4, v) \in \mathcal{S}_n$ as well. If we keep on doing so, there exist an integer $m(v,n) > 1$ and $u_o(v,n) \in \mathbb{Z}$ such that

$$\mathcal{U}_n(v) = \{u_o(v,n) + 4i : \ i = 1, \ldots, m(v,n)\}. \tag{4.10}$$

It is not hard to see that, for every $n > n_1$ and $v \in \mathcal{V}_n$, the integer $m(v,n)$ satisfies

$$\frac{2c\sqrt{n}}{4} - 2 < m(v,n) < \frac{2c\sqrt{n}}{4} + 2. \tag{4.11}$$

Condition 2). Recall that $(u,v) = (t,u,r)$. Let $U_{(v)}$ denote a random variable distributed as U given $V = v$, namely for every $z \in \mathbb{Z}$ it holds

$$P(U_{(v)} = z) = P(U = z | V = v).$$

For every $n > n_1$ and $v \in \mathcal{V}_n$, let us define:

$$p_n(i) := P\big(U_{(v)} = u_o(v,n) + 4i \,|\, U_{(v)} \in \mathcal{U}_n\big), \quad i = 1, \ldots, m(v,n).$$

The following lemma shows that the ratio $p_n(i+1)/p_n(i)$ tends to one with speed $O(n^{-\frac{1}{2}})$. The proof, given in the Appendix, is heavily based on the following well-known inequalities:

$$\begin{aligned} -\frac{3x}{2} &\leq \ln(1-x), \quad \text{for } 0 < x \leq 0.5 \\ \ln(1+x) &\leq x, \quad \text{for } x > -1. \end{aligned} \tag{4.12}$$

Lemma 4.3 *There exists an universal constant $K < \infty$ and $n_2 > n_1$ such that for every $n > n_2$ and $v \in \mathcal{V}_n$ it holds,*

$$1 - \frac{K}{\sqrt{n}} \leq \frac{p_n(i+1)}{p_n(i)} \leq 1 + \frac{K}{\sqrt{n}}, \quad i = 1, \ldots, m(v,n). \tag{4.13}$$

Recall (2.7), 2. and 4. Remark after Theorem 2.1. We are now ready to prove that the (conditional) variance of U increases linearly, i.e. condition **2)**:

Lemma 4.4 *There exist an universal constant $d > 0$ and $n_3 > n_2$ so that for every $n > n_3$ and for every $v \in \mathcal{V}_n$ it holds,*

$$\mathrm{Var}[U | U \in \mathcal{U}_n, V = v] = \mathrm{Var}[U_{(v)} | U_{(v)} \in \mathcal{U}_n] = \mathrm{Var}[U_{(v)} | U_{(v)} \in \mathcal{U}_n(v)] \geq dn. \tag{4.14}$$

Proof. Let $n > n_2$ and $v \in \mathcal{V}_n$. Recall (4.10) and (4.11). There exist constants $0 < d_1 < d_2$ so that for every n big enough, say $n > n_3 > n_2$, it holds

$$d_2\sqrt{n} < \frac{2c\sqrt{n}}{4} - 6 < \frac{2c\sqrt{n}}{4} + 2 < d_2\sqrt{n}.$$

From (4.11), it follows that for every $n > n_3$ and $v \in V$

$$d_1\sqrt{n} < m - 4 < m < d_2\sqrt{n}. \tag{4.15}$$

Take $n > n_3$. From (4.13), it follows that

$$1 - \frac{K}{\sqrt{n}} \leq \frac{p_n(i+1)}{p_n(i)} \leq 1 + \frac{K}{\sqrt{n}}$$

for $i = 1, \dots m-1$. Hence, for every $i, k \in \mathbb{N}$ such that $i + k \leq m$, it holds

$$\left(1 - \frac{K}{\sqrt{n}}\right)^k \leq \frac{p_n(i+k)}{p_n(i)} = \frac{p_n(i+1)}{p_n(i)} \cdot \frac{p_n(i+2)}{p_n(i+1)} \cdots \frac{p_n(i+k)}{p_n(i+k-1)} \leq \left(1 + \frac{K}{\sqrt{n}}\right)^k.$$

Recall (4.12), so that from the last inequality we get

$$\exp\left(-\frac{3K}{2\sqrt{n}}k\right) \leq \frac{p_n(i+k)}{p_n(i)} \leq \exp\left(\frac{K}{\sqrt{n}}k\right). \tag{4.16}$$

Thus, for every $1 \leq i, j \leq m$, we have

$$\frac{p_n(i)}{p_n(j)} \leq \exp\left(\frac{3K}{2\sqrt{n}}|i-j|\right) \leq \exp\left(\frac{3K}{2\sqrt{n}}m\right) \leq \exp\left(\frac{3K}{2}d_2\right) =: E. \tag{4.17}$$

From (4.17), it follows that $p_n(i) \leq (\min_i p_n(i))E$, so that

$$1 = \sum_{i=1}^{m} p_n(i) \leq m(\min_i p_n(i))E$$

and we obtain that for every $i = 1, \dots, m$, it holds

$$p_n(i) \geq (\min_i p_n(i)) \geq \frac{1}{mE} \geq \frac{1}{Ed_2\sqrt{n}}.$$

Now, the variance can be estimated as follows. Let $\bar{u} := E[U_{(v)}|U_{(v)} \in \mathcal{U}_n]$. Without loss of generality, let us assume that $\bar{u}(v,n) \leq u_o + 4(\frac{m+1}{2}) = u_o + 2m + 2$. Then for every $i \geq \frac{3m}{4}$, it holds that $|u_o + 4i - \bar{u}| = u_o + 4i - \bar{u} \geq m - 2$. Then

$$\begin{aligned}\mathrm{Var}[U_{(v)}|U_{(v)} \in \mathcal{U}_n] &= \sum_{i=1}^{m}(u_o + 4i - \bar{u})^2 p_n(i) \geq \sum_{i \geq \frac{3m}{4}}^{m}(u_o + 4i - \bar{u})^2 p_n(i) \\ &\geq (\frac{m}{4} - 1)(m-2)^2 \frac{1}{Ed_2\sqrt{n}} \geq \frac{d_1^3}{4Ed_2}n.\end{aligned}$$

■

Condition 1). The strategy is to look for a random mapping which satisfies assumptions **A1, A2** and **A3**, so that we check condition **1)** by applying Theorem 2.2. Recall that to apply Theorem 2.2, we additionally need that every point in the set $S_n \cap (\mathcal{U}_n \times \mathcal{V}_n)$ has to have sufficiently big probability so that the condition (2.17) is fulfilled. The following lemma, also proven in the Appendix, shows that the defined sets $\mathcal{U}_n$ and $\mathcal{V}_n$ indeed satisfy this additional condition:

Lemma 4.5 *There exist an universal constant $a > 0$ and $n_4 > n_3$ such that for any $n > n_4$ and $(u, v) \in \mathcal{S}_n \cap (\mathcal{U}_n \times \mathcal{V}_n)$ it holds*

$$P(U = u, V = v) \geq \frac{1}{an}.$$

Let us finally introduce the random transformation $\mathcal{R}$: Take $z = (x, y) \in \mathcal{X}_n \times \mathcal{X}_n$. Then in x, we choose uniformly at random a block of length $l - 1$ (among all the $b_{l-1}(x) \geq 1$ available blocks of length $l - 1$) and turn it to a block of length l. At the same time and independent from our previous choice, we choose uniformly at random a block of length $l + 1$ (among all the $b_{l+1}(x) \geq 1$ available blocks of length $l + 1$) and also turn it to a block of length l. We do not perform any change in y. Following our initial convention, $\tilde{z} := \mathcal{R}(z) = (\tilde{x}, y) \in \mathcal{X}_n \times \mathcal{X}_n$ is the sequence after applying this transformation.

Example 4.3 *As in a previous example with $l = 3$ and $n = 13$, let us take $x =$ 0011100111100 such that $b_2(x) = 2, b_3(x) = 1$ and $b_4(x) = 1$, with a rest at the end of length $r = 2$. In x, there are only two blocks of length $l - 1 = 2$ to pick from, each with probability $1/2$, namely 00 (most left one) or 00 (following most left one), and only one block of length $l + 1 = 4$ to pick from, with probability 1, namely 1111. Let us suppose that $\mathcal{R}$ picks up 00 (most left one) and 1111, then $\tilde{x}$ will look like this $\tilde{x} =$ 0001110011100.*

Note naturally that $b_l(\tilde{x}) = b_l(x) + 2$, $b_{l-1}(\tilde{x}) = b_{l-1}(x) - 1$ and $b_{l+1}(\tilde{x}) = b_{l+1}(x) - 1$, so that for $k_o = 4$ the condition (2.15) is satisfied.

We will prove that $\mathcal{R}$ satisfies assumptions **A3** and **A2**, but we do not prove in this paper that $\mathcal{R}$ satisfies assumption **A1**, because it would be too long (see 2. Remark after (2.16)) and the proof deals with another issues of random sequences comparison, which are different from the fluctuations ones we try to be focused on along the present article. It means that our main result, i.e. Theorem 4.1, delivers the linear fluctuations result assuming that **A1** is fulfilled. This is not restrictive, since in [37] assumption **A1** was already proven for the special case $q_1 = q_2 = q_3 = 1/3$ (as well as the linear fluctuations result). For the sake of completness, we include here the before mentioned result with our current notation:

Lemma 4.6 *Let $q_1 = q_2 = q_3 = 1/3$. There exist $n_0 < \infty$ and a constant $\alpha \in (0, 1)$ not depending on n but on l, such that for every $n > n_0$ the event*

$$E[L(\tilde{Z}) - L(Z)|Z] \geq \epsilon$$

happens with probability bigger than $1 - \exp[-n^{\alpha}]$.

Remark 4.1 (for readers who want to dig in the details of [37]) *As we have mentioned, the LCS of two sequences might have associated many different alignments, which we call optimal alignments. Lemma 4.2.1 in [37] showed that the set of realizations of X and Y such that their optimal alignments leave out at most a proportion q_0 of blocks, where $q_0 > \frac{4}{9(l-1)}$, have probability exponentially close to one as $n \to \infty$. The entire chapter 4 in [37] is dedicated to prove results of this type for a set of realizations of X and Y and their optimal alignments. Then, in Lemma 4.7.1 in [37] it is shown Lemma 4.6 as in the following form: the set of realizations of X and Y such that their optimal alignments satisfy $E[L(\tilde{Z}) - L(Z)|Z] \geq \epsilon_1$ has probability exponentially close to one as $n \to \infty$, for an arbitrary $\epsilon_1 > 0$. It is important to note that in [37], all the extra conditions of Lemma 4.7.1 and the extra work through the chapter 4 and chapter 6 are devoted to relate the value of ϵ_1 with the smallest possible length of the blocks $l > 0$ in order to get a sharp result for the order of the fluctuations of $L(Z)$. This relation is obtained in terms of an optimization problem which can be explicitely solved for this particular 3-multinomial model where $q_1 = q_2 = q_3 = 1/3$.*
The authors are working on a separate article about how to generalize Lemma 4.6 to the case $q_1, q_2, q_3 \in (0,1)$ such that $q_1 + q_2 + q_3 = 1$ (namely, the present and more general 3-multinomial block model).

Assumption A3. Recall that **A3** presupposes that for any $(u,v) \in \mathcal{S}_n \cap (\mathcal{U}_n \times \mathcal{V}_n)$, $b_{l-1}(u,v) \geq 1$ and $b_{l+1}(u,v) \geq 1$, which follows from (4.9). The following lemma proves **A3** :

Lemma 4.7 *Let $(u,v) \in \mathcal{S}_n$ be such that $b_{l-1}(u,v) \geq 1$ and $b_{l+1}(u,v) \geq 1$. If $Z \sim P_{(u,v)}$, then $\tilde{Z} \sim P_{(u+4,v)}$.*

Proof. The random variables U, T and R are independent of Y. Hence, $P_{(u,v)} = P^x_{(u,v)} \times P^y$, where $P^x_{(u,v)}$ is the conditional distribution of X given $\{(U,V) = (u,v)\}$, P^y is the law of Y (actually $P^x = P^y$) and $\times$ stands for the product measure. Also the block transformation applies to X, only. Thus, it suffices to show that if $X \sim P^x_{(u,v)}$, then $\tilde{X} \sim P^x_{(u+4,v)}$. For proving this, we follow the approach in Lemma 3.3.
The first step of the proof is to explicitly compute an expression for $P^x_{(u,v)} = P^x_{(u,t,r)}$. By (4.7), we have that for any $x \in S(b_{l-1}(u,v), b_l(u,v), b_{l+1}(u,v))$ it holds:

$$\begin{aligned} P^x_{(u,v)}(x) &= P(X = x | U = u, T = t, R = r) \\ &= P(X = x | B_{l-1} = b_{l-1}(u,v), B_l = b_l(u,v), B_{l+1} = b_{l+1}(u,v)) \\ &= \frac{1}{2} \binom{t}{b_{l-1}(u,v)\ b_l(u,v)\ b_{l+1}(u,v)}^{-1}. \end{aligned} \tag{4.18}$$

The second step of the proof is to actually compute the distribution of $\tilde{X}$. For that, we need to investigate the effect of the block-transformation on the distribution of X. Let us fix $x \in \mathcal{X}$ and denote $b_1 := b_{l-1}(x), b_2 := b_l(x)$ and $b_3 := b_{l+1}(x)$, and (u,v) its

corresponding triple. Let us define $\mathcal{A}(x)$ the set of all strings that are possible outcomes after applying the block transformation to x, namely if $\tilde{x} \in \mathcal{A}(x)$ then necessarily

$$b_{l-1}(\tilde{x}) = b_1 - 1, b_l(\tilde{x}) = b_2 + 2, b_{l+1}(\tilde{x}) = b_3 - 1.$$

However, not every string $y \in \mathcal{X}_n$ such that $b_{l-1}(y) = b_1 - 1, b_l(y) = b_2 + 2$ and $b_{l+1}(y) = b_3 - 1$ belongs to $\mathcal{A}(x)$. By using (4.4), it is straightforward to see that triple $(b_{l-1}(\tilde{x}), b_l(\tilde{x}), b_{l+1}(\tilde{x}))$ corresponds to the triple $(u+4, v)$. Since the block-transformation picks up blocks uniformly, then after applying it to x, every element of $\mathcal{A}(x)$ has the same probability to occur. Formally,

$$P(\tilde{X} = \tilde{x}|X = x) = \begin{cases} \theta & \text{if } \tilde{x} \in \mathcal{A}(x) \\ 0 & \text{otherwise} \end{cases} \tag{4.19}$$

where $\theta \in (0, 1]$ is a constant which depends on x only through (b_1, b_2, b_3) (or equivalently through (u, v)). The block-transformation only changes blocks of length $l-1$ and $l+1$, so $\theta = 1/(b_1 \cdot b_3)$. The last ingredient is to define the set $\mathcal{B}(\tilde{x}) := \{x \in \mathcal{X} : \tilde{x} \in \mathcal{A}(x)\}$. The cardinality of this set is

$$|\mathcal{B}(\tilde{x})| = 2\binom{b_l(\tilde{x})}{2}$$

because, after the block-transformation, each block of length l could have eventually came from two previous shorter or longer blocks, so the $\binom{b_l(\tilde{x})}{2}$, and can be either of 1's or of 0's, so the 2.
To end the proof, let us consider $X \sim P_{(u,v)}$. Then, we get the corresponding triple

$$(b_{l-1}(u, v), b_l(u, v), b_{l+1}(u, v))$$

depending only on (u, v). To keep the notation light in what follows, let us call $b_1^* := b_{l-1}(u, v), b_2^* := b_l(u, t, v)$ and $b_3^* := b_{l+1}(u, v)$. We aim to find $P(\tilde{X} = \tilde{x}) = P(\tilde{X} = \tilde{x}|U = u, T = t, R = r)$. Note that for every $\tilde{x} \in \mathcal{X}_n$, there are two possibilities: either $\mathcal{B}(\tilde{x}) \cap S(u, v) = \emptyset$ or

$$\mathcal{B}(\tilde{x}) \subset S(u, v).$$

The second case holds if and only if

$$b_{l-1}(\tilde{x}) = b_1^* - 1, b_l(\tilde{x}) = b_2^* + 2, b_{l+1}(\tilde{x}) = b_3^* - 1$$

or, equivalently $\tilde{x} \in S(u+4, v)$. In this case, by (4.18) and (4.19) we have:

$$\begin{aligned}
P(\tilde{X} = \tilde{x} | U = u, T = t, R = r) &= \sum_{x \in \mathcal{B}(\tilde{x})} P(\tilde{X} = \tilde{x} | X = x) P(X = x | U = u, T = t, R = r) \\
&= \sum_{x \in \mathcal{B}(\tilde{x})} \frac{1}{b_1^* b_3^*} P(X = x | U = t, T = t, R = r) \\
&= \frac{1}{2 b_1^* b_3^*} \binom{t}{b_1^*\ b_2^*\ b_3^*}^{-1} |\mathcal{B}(\tilde{x})| \\
&= \frac{1}{2 b_1^* b_3^*} \binom{t}{b_1^*\ b_2^*\ b_3^*}^{-1} 2 \binom{b_2^* + 2}{2} \\
&= \frac{1}{2} \binom{t}{b_1^* - 1\ \ b_2^* + 2\ \ b_3^* - 1}^{-1} \\
&= P^x_{(u+4,v)}(\tilde{x}).
\end{aligned}$$

■

Assumption A2. This assumption means that in the worse case, the length of the LCS decreases in A units after the block-transformation. Let $z = (x, y)$ be a realization of Z, w_{l-1} be the block of length $l-1$ and w_{l+1} be the block of length $l+1$ in x that $\mathcal{R}$ has chosen. Note that the decrease in the length of the LCS comes from the following fact: if in every optimal alignment producing the LCS of x and y all the bits of w_{l+1} are aligned, then it is clear that deleting one bit of w_{l+1} will decrease the length of the LCS of x and y only by 1. Later, by adding a new bit in w_{l-1}, we cannot get an even lower length of the LCS of x and y (in the worst case, we stay the same). Therefore, $A = 1$.

Example 4.4 *Take* $l = 2$, $x = 11100010101101$ *and* $y = 11101100101000$. *Then* $L_{14}(x, y) = 10$ *could be represented by the alignment*

x	1	1	1	0	0	0	–	1	–	0	1	0	1	1	0	–	–	1	–
y	1	1	1	0	–	–	1	1	0	0	1	0	1	–	0	0	0	–	0
$L_{14}(x, y)$	1	1	1	0				1		0	1	0	1		0				

Suppose that $\mathcal{R}$ *deletes from the block* $w_3 = 111$ *(first block, from left to right) one symbol. The minimum gain for the length of an LCS is when* $\mathcal{R}$ *adds the extra symbol either to the fifth block in* x *of length one (from the left to the right):*

x	1	1	–	0	0	0	–	1	–	0	1	1	0	1	1	0	–	–	1	–
y	1	1	1	0	–	–	1	1	0	0	1	–	0	1	–	0	0	0	–	0
$L_{14}(\tilde{x}, y)$	1	1		0				1		0	1		0	1		0				

or to the sixth block in x *of length one (from the left to the right):*

x	1	1	–	0	0	0	–	1	–	0	1	0	0	1	1	0	–	–	1	–
y	1	1	1	0	–	–	1	1	0	0	1	–	0	1	–	0	0	0	–	0
$L_{14}(\tilde{x}, y)$	1	1		0				1		0	1		0	1		0				

In both cases, we get $L_{14}(\tilde{x}, y) = 9$.

We now state the main theorem of the section, about the linear fluctuations of the length of the LCS in the 3-multinomial block model:

Theorem 4.1 *Assume that the block-transformation satisfies* **A1**. *Then there exists an universal constant* $b > 0$ *and* $n_6 < \infty$ *so that for every* $n > n_6$, *it holds*

$$\mathrm{Var}[L(Z)] \geq b \cdot n. \tag{4.20}$$

Proof. Let us first check the assumptions of Theorem 2.2. **A1** holds by hypothesis; our block-transformation is such that **A2** holds with $A = 1$ (as discussed above). Let now n_4 be as in Lemma 4.5. Recall that $n_0 < n_1 < n_2 < n_3 < n_4 < \infty$. Hence, for any $n > n_4$, (4.9), (4.14) and (2.17) hold. Moreover, from Lemma 4.1, it holds $P(U \in \mathcal{U}_n, V \in \mathcal{V}_n) \geq 0.9$. The condition (4.9) states that for every $(u, v) \in \mathcal{S}_n \cap \{\mathcal{U}_n \times \mathcal{V}_n\}$, we have that $b_{l-1}(u, v) \geq 1$ and $b_{l+1}(u, v) \geq 1$. Hence, the block-transformation is possible, and Lemma 4.7 now establishes **A3**. Therefore, for every $n > n_4$, the assumptions of Theorem 2.2 are fulfilled and, therefore, there exists $n_5 > n_4$ so that for for every $n > n_5$, the assumptions of Theorem 2.1 hold with $\delta = \frac{\epsilon_0}{2}$.
We now apply Theorem 2.1. As just showed, the assumption **1**) holds for any $n > n_5$; as explained at the beginning of Subsection 4.2, the assumption **3)** holds with $k_o = 4$. By (4.14), $\psi_n = dn$. By Theorem 2.1, thus, the lower bound (4.20) exists with

$$b = \frac{9\epsilon_o^2 d}{640}.$$

■

Acknowledgments

F.T. would like to thank the Estonian Science Foundation through the Grant nr. 9288 and targeted financing project SF0180015s12 for making possible a two weeks research stay at Tartu University visiting J.L. while working in the core of this article, as well as the DFG through the SFB 878 at University of Münster for financial support while the research stay. J.L would like to thank the Estonian Science Foundation through the Grant nr. 9288 and targeted financing project SF0180015s12 for supporting a short research stay at University of Münster while the finishing of this article.

A Appendix

Proposition A.1 *Given* $\epsilon > 0$ *there exist a constant* $c'(\epsilon)$ *not depending on* n *but on* ϵ *and* $n_0(\epsilon) < \infty$ *such that for every* $n > n_0$

$$P\left(\left|\frac{B_{l-1} - q_1\frac{n}{\mu}}{\sqrt{n}}\right| \leq c', \quad \left|\frac{B_l - q_2\frac{n}{\mu}}{\sqrt{n}}\right| \leq c', \quad \left|\frac{B_{l+1} - q_3\frac{n}{\mu}}{\sqrt{n}}\right| \leq c'\right) \geq 1 - \epsilon. \tag{A.1}$$

Proof. It suffices to show that for every $\epsilon > 0$ there exists $\gamma_i(\epsilon) > 0$ $i = 1, 2, 3$ so that

$$P\left(|\frac{B_{l_i} - q_i \frac{n}{l}}{\sqrt{n}}| \leq \gamma_i\right) \geq 1 - \epsilon, \tag{A.2}$$

for $i = 1, 2, 3$ and $l_1 := l - 1$, $l_2 := l$ and $l_3 := l + 1$. From (A.2) the bound (A.1) trivially follows by taking

$$c'(\epsilon) := \max_i \left\{\gamma_1(\frac{\epsilon}{3}), \gamma_2(\frac{\epsilon}{3}), \gamma_3(\frac{\epsilon}{3})\right\}.$$

Even more, we shall only show the existence of one-sided bound for B_l: for every ϵ, there exists $\gamma(\epsilon)$ so that

$$P\left(\frac{B_l - q_2 \frac{n}{\mu}}{\sqrt{n}} \leq \gamma(\epsilon)\right) \geq 1 - \epsilon. \tag{A.3}$$

The other side follows from the same arguments. Since in this proof, we consider q_2, only, let for that proof $q := q_2$. Let α, β, γ be positive real numbers and $m \in \mathbb{N}$. We define the random variables ξ_i and events $A(\alpha, m), B(\beta, n)$ and $C(\gamma, n)$ as following:

$$\begin{aligned}
\xi_i &:= \begin{cases} 1 & \text{if } W_i = l \\ 0 & \text{otherwise} \end{cases} \\
A(\alpha, m) &:= \left\{\xi_1 + \cdots + \xi_m \leq qm + \alpha\sqrt{m}\right\} \\
B(\beta, n) &:= \left\{B_{l-1} + B_l + B_{l+1} \leq \frac{n}{\mu} + \beta\sqrt{n}\right\} \\
C(\gamma, n) &:= \left\{B_l \leq 3\frac{n}{\mu} + \gamma\sqrt{n}\right\}.
\end{aligned}$$

Note that for any m, the event $B_{l-1} + B_l + B_{l+1} > m$ implies that in the sequence X, there are more than m blocks. That, in turn, means that the m first blocks cover less than n bits of $X_1, X_2, \ldots$ or, equivalently, $W_1 + \cdots + W_m \leq n$. Let

$$m(n, \beta) := \frac{n}{\mu} + \beta\sqrt{n}.$$

Note that $\frac{n}{m} - \mu < 0$. Thus

$$\begin{aligned}
P(B^c(\beta, n)) &\leq P\left(W_1 + \cdots + W_m \leq n\right) \\
&= P\left(\frac{W_1 + \cdots + W_m}{m} - \mu \leq \frac{n}{m} - \mu\right) \\
\text{(by (2.2) with P}(|W_1 - \mu| \leq 2) = 1) &\leq \exp\left(-\frac{m}{8}\left(\frac{n}{m} - \mu\right)^2\right) \\
&= \exp\left(-\frac{\mu^3\beta^2}{8}\left(\frac{1}{1 + \frac{\beta\mu}{\sqrt{n}}}\right)\right).
\end{aligned} \tag{A.4}$$

Now, for any $\epsilon > 0$, we can find $\beta_o = \beta(\epsilon)$ so big that $\exp[-\frac{\mu^3\beta_o^2}{2 \cdot 8}] < \frac{\epsilon}{2}$. An then, one can take $n_0(\epsilon)$ so big that $\frac{\mu\beta_o}{\sqrt{n_o}} < 1$. Hence, for any $n > n_o$,

$$P(B^c(n, \beta_o)) < \frac{\epsilon}{2}. \tag{A.5}$$

Let

$$\alpha_0 := \sqrt{\frac{2q(1-q)}{\epsilon}}.$$

Then by (2.1), for any m

$$P(A^c(\alpha_0, m)) = P\left(\frac{\xi_1 + \cdots + \xi_m}{m} - q \geq \frac{\alpha_0}{\sqrt{m}}\right) \leq \frac{\epsilon}{2}. \tag{A.6}$$

Finally, if we define

$$\gamma(\beta, \alpha_0) := q\beta + \alpha_0\sqrt{\frac{1}{\mu} + \beta}$$

then we have

$$\left.\begin{array}{c} \xi_1 + \cdots + \xi_{m(\beta,n)} \leq qm(\beta,n) + \alpha_0\sqrt{m(\beta,n)} \\ B_{l-1} + B_l + B_{l-1} \leq m(\beta,n) \end{array}\right\} \Rightarrow B_l \leq qm(\beta,n) + \alpha_0\sqrt{m(\beta,n)} \leq q\frac{n}{\mu} + \gamma\sqrt{n}.$$

Therefore

$$A(\alpha_0, m) \cap B(\beta_o, n) \subseteq C(\gamma, n). \tag{A.7}$$

Now take $n > n_0$, $m_0 := m(\beta_0, n)$ and $\gamma_0 := \gamma(\beta_0, \alpha_0)$ and use (A.6), (A.4) and (A.7) to get

$$P\left(\frac{B_l - q\frac{n}{\mu}}{\sqrt{n}} > \gamma_0\right) = P(C^c(\gamma_0, n)) \leq P(A^c(\alpha_0, m_0)) + P(B^c(\beta_0, n)) \leq \epsilon. \tag{A.8}$$

■

Proof of Lemma 4.1. By Proposition A.1, for any $\epsilon > 0$ there exist $c' = c'(\epsilon) > 0$ and $n_0(\epsilon) < \infty$ such that:

$$\begin{aligned} P\left(\left|\frac{B_{l-1} + B_l + B_{l+1} - \frac{n}{\mu}}{\sqrt{n}}\right| > 3c'\right) &\leq P\left(\left|\frac{B_{l-1} - q_1\frac{n}{\mu}}{\sqrt{n}}\right| + \left|\frac{B_l - q_2\frac{n}{\mu}}{\sqrt{n}}\right| + \left|\frac{B_{l+1} - q_3\frac{n}{\mu}}{\sqrt{n}}\right| > 3c'\right) \\ &\leq \sum_{i=1} P\left(\left|\frac{B_{l_i} - q_i\frac{n}{\mu}}{\sqrt{n}}\right| > c'\right) \leq 3\epsilon \\ P\left(\left|\frac{B_l - B_{l-1} - B_{l+1} - \frac{n}{\mu}(q_2 - q_1 - q_3)}{\sqrt{n}}\right| > 3c'\right) &\leq P\left(\left|\frac{B_{l-1} - q_1\frac{n}{\mu}}{\sqrt{n}}\right| + \left|\frac{B_l - q_2\frac{n}{\mu}}{\sqrt{n}}\right| + \left|\frac{B_{l+1} - q_3\frac{n}{\mu}}{\sqrt{n}}\right| > 3c'\right) \\ &\leq \sum_{i=1} P\left(\left|\frac{B_{l_i} - q_i\frac{n}{\mu}}{\sqrt{n}}\right| > c'\right) \leq 3\epsilon \end{aligned}$$

for any $n > n_0$, where as before $l_1 = l-1, l_2 = l$ and $l_3 = l+1$. Then, it directly follows:

$$P\big(U \notin \mathcal{U}_n^{3c'}, V \notin \mathcal{V}_n^{3c'}\big) \geq 6\epsilon$$

from where the proof is completed by taking $c := 3c'$ and $\epsilon = 1/60$.

■

Proof of Lemma 4.2. Let $n > n_0$. Take $(u, v) \in \mathcal{S}_n \cap (\mathcal{U}_n \times \mathcal{V}_n)$ (By lemma 4.1, the set $\mathcal{S}_n \cap (\mathcal{U}_n \times \mathcal{V}_n)$ is not empty). In particular,

$$t = \frac{n}{\mu} + \sigma_1, \quad u = \frac{n}{\mu}(q_2 - q_1 - q_3) + \sigma_2, \quad r \in \{1, \ldots, l+1\},$$

for $\sigma_1, \sigma_2 \in [-c\sqrt{n}, c\sqrt{n}\,]$. Let us start by showing that $b_{l_i}(u, t, r) \geq 1$. Recall $\mu = l + q_3 - q_1$. From (4.4) and $1 \leq r < l + 1$ we get

$$\begin{aligned} b_{l-1}(u,t,r) &= q_1\frac{n}{\mu} + \frac{2\sigma_1 l + (\sigma_1 - \sigma_2) + 2r}{4} \\ &\geq q_1\frac{n}{\mu} - \frac{c(l+1)}{2}\sqrt{n} \geq 1 \end{aligned}$$

provided $n > n_{1,1}(c)$. Also, by using (4.4) and $1 < r < l + 1$ we get

$$b_l(u,t,r) = q_2\frac{n}{\mu} + \frac{\sigma_1 + \sigma_2}{2} \quad \geq \quad q_2\frac{n}{\mu} - c\sqrt{n} \geq 1$$

provided $n > n_{1,2}(c)$. Finally, also by using (4.4) and $1 \leq r < l + 1$ we get

$$\begin{aligned} b_{l+1}(u,t,r) &= q_3\frac{n}{\mu} - \frac{2\sigma_1 l + (\sigma_2 - \sigma_1) + 2r}{4} \\ &\geq q_3\frac{n}{\mu} - \frac{c(l+1)}{2}\sqrt{n} - \frac{l-2}{2} \geq 1 \end{aligned}$$

provided $n > n_{1,3}(c)$. So, for having simultaneously the three lower bounds we need to take $n > n_1^1 := \max\{n_{1,1}, n_{1,2}, n_{1,3}\}$.
For the absolute value bounds, we proceed in the same way:

$$\begin{aligned} \left|b_{l-1}(u,t,r) - q_1\frac{n}{\mu}\right| = \left|\frac{2\sigma_1 l + (\sigma_1 - \sigma_2) + 2r}{4}\right| &\leq \frac{c(l+2)}{2}\sqrt{n}, \quad \text{for } n \geq (l-2)^2/c^2 \\ \left|b_l(u,t,r) - q_2\frac{n}{\mu}\right| = \left|\frac{\sigma_1 + \sigma_2}{2}\right| &\leq c\sqrt{n}, \quad \text{for } n \geq 1 \\ \left|b_{l+1}(u,t,r) - q_3\frac{n}{\mu}\right| = \left|-\frac{2\sigma_1 l + (\sigma_2 - \sigma_1) + 2r}{4}\right| &\leq \frac{c(l+1)}{2}\sqrt{n}, \quad \text{for } n \geq (l-2)^2/c^2. \end{aligned}$$

Thus, all the above three upper bounds hold for $n_1^2 := \max\{1, (l-2)/c\}$. In order to obtain (4.9), it is enough to take $n_1 = \max\{n_1^1, n_1^2\}$ and the universal constant $\alpha := c(l+2)/2$ which does not depend on (u, t, r). Without loss of generality, we can take $n_1 > n_0$.

■

Proof of Lemma 4.3. Let $n > n_1$ and consider $(u, v) \in \mathcal{S}_n \cap (\mathcal{U}_n \times \mathcal{V}_n)$. Let

$$(b_{l-1}(u,v), b_l(u,v), b_{l+1}(u,v))$$

be the solution of (4.4). We have already seen that

$$\begin{aligned} b_{l-1}(u+4,v) &= b_{l-1}(u,v)-1 \\ b_l(u+4,v) &= b_l(u,v)+2 \\ b_{l+1}(u+4,v) &= b_{l+1}(u,v)-1. \end{aligned}$$

Therefore, by using these relations and (4.8) (formula of the multinomial distribution for (U,T,R)) we get:

$$\frac{P(U_{(v)}=u+4|U_{(v)}\in\mathcal{U}_n)}{P(U_{(v)}=u|U_{(v)}\in\mathcal{U}_n)} = \frac{b_{l-1}(u,v)b_{l+1}(u,v)}{(b_l(u,v)+1)(b_l(u,v)+2)}\cdot\frac{q_2^2}{q_1q_3}. \tag{A.9}$$

By using Lemma 4.2, there exist an universal constant $\alpha>0$ such that:

$$\frac{b_{l-1}(u,v)b_{l+1}(u,v)}{(b_l(u,v)+1)(b_l(u,v)+2)}\cdot\frac{q_2^2}{q_1q_3} \geq \frac{\left(q_1\frac{n}{\mu}-\alpha\sqrt{n}\right)\left(q_3\frac{n}{\mu}-\alpha\sqrt{n}\right)}{\left(1+q_2\frac{n}{\mu}+\alpha\sqrt{n}\right)\left(2+q_2\frac{n}{\mu}+\alpha\sqrt{n}\right)}\cdot\frac{q_2^2}{q_1q_3} \tag{A.10}$$

$$\frac{b_{l-1}(u,v)b_{l+1}(u,v)}{(b_l(u,v)+1)(b_l(u,v)+2)}\cdot\frac{q_2^2}{q_1q_3} \leq \frac{\left(q_1\frac{n}{\mu}+\alpha\sqrt{n}\right)\left(q_3\frac{n}{\mu}+\alpha\sqrt{n}\right)}{\left(1+q_2\frac{n}{\mu}-\alpha\sqrt{n}\right)\left(2+q_2\frac{n}{\mu}-\alpha\sqrt{n}\right)}\cdot\frac{q_2^2}{q_1q_3} \tag{A.11}$$

for every $n>n_1$. Recall the inequalities (4.12). Let us start by looking at (A.10):

$$\begin{aligned} \frac{\left(q_1\frac{n}{\mu}-\alpha\sqrt{n}\right)\left(q_3\frac{n}{\mu}-\alpha\sqrt{n}\right)}{\left(1+q_2\frac{n}{\mu}+\alpha\sqrt{n}\right)\left(2+q_2\frac{n}{\mu}+\alpha\sqrt{n}\right)}\cdot\frac{q_2^2}{q_1q_3} &\geq \frac{\left(1-\frac{\mu\alpha/q_1}{\sqrt{n}}\right)\left(1-\frac{\mu\alpha/q_3}{\sqrt{n}}\right)}{\left(1+\frac{2\mu\alpha/q_2}{\sqrt{n}}\right)^2} \\ &= \exp\left[\ln\left(1-\frac{\mu\alpha/q_1}{\sqrt{n}}\right)+\ln\left(1-\frac{\mu\alpha/q_3}{\sqrt{n}}\right)\right. \\ &\quad \left.-2\ln\left(1+\frac{2\mu\alpha/q_2}{\sqrt{n}}\right)\right] \\ \left(\text{from (4.12) for every } n>\max\{\tfrac{\mu^2\alpha^2}{q_1^2},\tfrac{\mu^2\alpha^2}{q_3^2}\}\right) &\geq \exp\left[-\frac{\mu\alpha}{2\sqrt{n}}\left(\frac{3}{q_1}+\frac{8}{q_2}+\frac{3}{q_3}\right)\right] \\ &\geq 1-\frac{\mu\alpha}{2}\left(\frac{3}{q_1}+\frac{8}{q_2}+\frac{3}{q_3}\right)\frac{1}{\sqrt{n}}. \end{aligned} \tag{A.12}$$

Next, let us fix an arbitrary $\epsilon>0$. Then, there exists $n_{2,1}<\infty$ such that the rest of the Taylor's expansion of the function $f(x)=e^{-x}$ at $\xi:=\frac{\mu\alpha}{\sqrt{n}}\left(\frac{1}{q_1}+\frac{6}{q_2}+\frac{1}{q_3}\right)$ satisfies:

$$R(\xi):=\left|\frac{f''(\xi)}{2}\right|\xi^2=\frac{\mu^2\alpha^2}{n}\left(\frac{1}{q_1}+\frac{6}{q_2}+\frac{1}{q_3}\right)^2\exp\left(-\frac{\mu\alpha}{\sqrt{n}}\left(\frac{1}{q_1}+\frac{6}{q_2}+\frac{1}{q_3}\right)\right)\leq\epsilon \tag{A.13}$$

for every $n > n_{2,1}$. Note that $n_{2,1}$ does not depend on (u, v) but only on known fixed constants. Now, let us look at (A.11):

$$\begin{aligned}
\frac{\left(q_1\frac{n}{\mu}+\alpha\sqrt{n}\right)\left(q_3\frac{n}{\mu}+\alpha\sqrt{n}\right)}{\left(1+q_2\frac{n}{\mu}-\alpha\sqrt{n}\right)\left(2+q_2\frac{n}{\mu}-\alpha\sqrt{n}\right)}\cdot\frac{q_2^2}{q_1q_3} &\leq \frac{\left(1+\frac{\mu\alpha/q_1}{\sqrt{n}}\right)\left(1+\frac{\mu\alpha/q_3}{\sqrt{n}}\right)}{\left(1-\frac{2\mu\alpha/q_2}{\sqrt{n}}\right)^2} \\
&= \exp\left[\ln\left(1+\frac{\mu\alpha/q_1}{\sqrt{n}}\right)+\ln\left(1+\frac{\mu\alpha/q_3}{\sqrt{n}}\right)\right. \\
&\quad \left.-2\ln\left(1-\frac{2\mu\alpha/q_2}{\sqrt{n}}\right)\right] \\
\left(\text{from (4.12) for every } n>\tfrac{16\mu^2\alpha^2}{q_2^2}\right) &\leq \exp\left[\frac{\mu\alpha}{\sqrt{n}}\left(\frac{1}{q_1}+\frac{6}{q_2}+\frac{1}{q_3}\right)\right] \\
&\leq 1+\mu\alpha\left(\frac{1}{q_1}+\frac{6}{q_2}+\frac{1}{q_3}\right)\frac{1}{\sqrt{n}}+|R(\xi)| \\
\text{from (A.13)} &\leq 1+\mu\alpha\left(\frac{1}{q_1}+\frac{6}{q_2}+\frac{1}{q_3}\right)\frac{1}{\sqrt{n}}+\epsilon \qquad \text{(A.14)}
\end{aligned}$$

for every $n > n_{2,2} := \max\{n_{2,1}, \frac{16\mu^2\alpha^2}{q_2^2}\}$. Finally, from (A.12) and (A.14) we have:

$$1-\frac{\mu\alpha}{2}\left(\frac{3}{q_1}+\frac{8}{q_2}+\frac{3}{q_3}\right)\frac{1}{\sqrt{n}} \leq \frac{P(U_{(v)}=u+4|U_{(v)}\in\mathcal{U}_n)}{P(U_{(v)}=u|U_{(v)}\in\mathcal{U}_n)} \leq 1+\mu\alpha\left(\frac{1}{q_1}+\frac{6}{q_2}+\frac{1}{q_3}\right)\frac{1}{\sqrt{n}}+\epsilon$$

for any arbitrary $\epsilon > 0$. ¿From this last inequality, we can find a constant $K > 0$ not depending on n neither on (u, v) and n_2 bigger than $n_{2,2}$ and n_1 such that (4.13) holds for every $n > n_2$.

■

Proof of Lemma 4.5. The proof is based on the Corollary 2.2. Define

$$\beta := (\alpha + c)\sqrt{2\mu}, \qquad \text{(A.15)}$$

where α is as in (4.9) and choose $n_4 > n_3$ so big that simultaneously $\frac{n_4}{\mu} - c\sqrt{n} > m_o(\beta)$ and $\sqrt{n_4} \geq 2c\mu$. Here $m_o(\beta)$ is as in Corollary 2.2. From these inequalities, it follows that whenever $n > n_4$, then

$$\frac{n}{\mu} - c\sqrt{n} \geq \max\left\{\frac{1}{2\mu}n, m_o\right\}. \qquad \text{(A.16)}$$

Take now $n > n_4$ and $(u, t, r) \in \mathcal{S}_n \cap (\mathcal{U}_n \times \mathcal{V}_n)$. By (A.16),

$$t \geq \frac{n}{\mu} - c\sqrt{n} > m_o(\beta), \quad 2\mu t \geq n. \qquad \text{(A.17)}$$

Use now (4.9) and the definition of $\mathcal{V}_n$ to see that for every $i = 1, 2, 3$,

$$|b_{l_i} - tq_i| \leq |b_{l_i} - q_i(\frac{n}{\mu})| + q_i|\frac{n}{\mu} - t| \leq \alpha\sqrt{n} + q_i c\sqrt{n} \leq (\alpha + c)\sqrt{n} \leq (\alpha + c)\sqrt{2\mu}\sqrt{t},$$

where the last inequality follows from (A.17) and $b_{l_i} = b_{l_i}(u,t,r)$. Apply (2.5) with β as in (A.15), $m = t$, $p_1 = q_1$, $p_2 = q_2$, $p_3 = q_3$ and $i = b_{l_1}$, $j = b_{l_2}$. Then by (4.8)

$$P(U = u, T = t, R = r) = \binom{t}{b_{l_1}\ b_{l_2}\ b_{l_3}} q_1^{b_{l_1}} q_2^{b_{l_2}} q_3^{b_{l_3}} p(r) \geq \frac{p(r)}{b(\beta)n} \geq \frac{q_3}{b(\beta)n}, \tag{A.18}$$

where the last inequality comes from the fact that $p(r) \geq q_3$. Thus Lemma 4.5 is proven with

$$a = \frac{b(\beta)}{q_3}.$$

■

References

[1] Kenneth S. Alexander. *The rate of convergence of the mean length of the longest common subsequence.* Ann. Appl. Probab. **4**(4): 1074–1082, 1994.

[2] Kenneth S. Alexander. *Approximation of subadditive functions and convergence rates in limiting-shape results.* Ann. Appl. Probab., **25**(1):30–55, 1997.

[3] R.A. Baeza-Yates, R. Gavaldà, G. Navarro, and R. Scheihing. *Bounding the expected length of longest common subsequences and forests.* Theory Comput. Syst., **32**(4):435–452, 1999.

[4] Andrew Barron, Lucien Birgé, and Pascal Massart. *Risk bounds for model selection via penalization.* Probab. Theory Related Fields, **113**(3):301–413, 1999.

[5] H. S. Booth, S.F. MacNamara, O.M. Nielsen, and S.R. Wilson. *An iterative approach to determining the length of the longest common subsequence of two strings.* Methodology Comput. Appl. Probab., **6**(4):401–421, 2004.

[6] Federico Bonetto and Heinrich Matzinger. *Fluctuations of the longest common subsequence in the case of 2- and 3-letter alphabets.* Latin American J. Proba. Math. **2**: 195–216, 2006.

[7] J. Boutet de Monvel *Extensive simulations for longest common subsequences* Eur. Phys. J. B **7**: 293–308, 1999.

[8] Clement Durringer, Jüri Lember and Heinrich Matzinger. *Deviation from the mean in sequence comparison with a periodic sequence.* ALEA. **3**: 1–29, 2007.

[9] N. Christianini and M. W. Hahn. *Introduction to Computational Genomics.* Cambridge University Press, 2007.

[10] Václáv Chvatal and David Sankoff. *Longest common subsequences of two random sequences.* J. Appl. Probab. **12**: 306–315, 1975.

[11] Vlado Dancik. *Expected length of longest common subsequences.* PhD dissertation, Department of Computer Science, University of Warwick, 1994.

[12] Vlado Dancik and Mike Paterson. *Upper bounds for the expected length of a longest common subsequence of two binary sequences.* Random Structures Algorithms, **6**(4):449–458, 1995.

[13] Joseph G. Deken. *Some limit results for longest common subsequences.* Discrete Math., **26**(1):17–31, 1979.

[14] Luc Devroye, Gabor Lugosi, and Laszlo Gyorfi. *A probabilistic theory of pattern recognition.* Springer-Verlag, New York, 1996.

[15] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998.

[16] J. Fu and W. Lou. *Distribution of the length of the longest common subsequence of two multi-state biological sequences.* Journal of Statistical Planning and Inference, **138**:3605–3615, 2008.

[17] Geoffrey Grimmett and David Stirzaker *Probability and Random Processes* Oxford University Press, 2001. Third Edition.

[18] Rick Durrett *Probability: Theory and Examples* Thomson, 2005. Third Edition.

[19] Raphael Hauser, Heinrich Matzinger and Clement Durringer. *Approximation to the mean curve in the lcs-problem.* Stochastic Proc. and Appl., **118**(4):629–648, 2008.

[20] Christian Houdre and Heinrich Matzinger. *Fluctuations of the Optimal Alignment Score with and Asymmetric Scoring Function.* [arXiv:math/0702036]

[21] Chirstian Houdre and Jüri Lember and Heinrich Matzinger. *On the longest common increasing binary subsequence.* C.R. Acad. Sci. Paris, Ser. **I 343**:589–594, 2006.

[22] Marcos A. Kiwi, Martin Loebl, and Jirí Matousek. *Expected length of the longest common subsequence for large alphabets.* Advances in Mathematics, **197**(2):480–498, 2005.

[23] Raphael Hauser and Servet Martinez and Heinrich Matzinger *Large deviation based upper bounds for the LCS-problem* Advances in Applied Probability. Volume 38: 827-852, 2006.

[24] Jüri Lember and Heinrich Matzinger *Standard Deviation of the Longest Common Subsequence.* Ann. Probab. **37**(3): 1192–1235, 2009.

[25] Jüri Lember, Heinrich Matzinger and Felipe Torres. *The rate of the convergence of the mean score in random sequence comparison.* Ann. Appl. Probab. **22**(3): 1046–1058, 2012.

[26] Jüri Lember, Heinrich Matzinger and Felipe Torres. *Proportion of gaps and fluctuations of the optimal score in random sequence comparison.* to appear in Proceedings in Honour of Friedrich Götze's 60th-birthday. Springer, 2012

[27] Chin-Yew Lin and Franz Josef Och. *Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics.* In ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, page 605, 2004.

[28] George S. Lueker. *Improved bounds on the average length of longest common subsequences.* J. ACM, **56**(3):1–38, 2009.

[29] Heinrich Matzinger and Felipe Torres. *Fluctuation of the longest common subsequence for sequences of independent blocks.* [arvix math.PR/1001.1273v3]

[30] Heinrich Matzinger and Felipe Torres. *Random modification effect in the size of the fluctuation of the LCS of two sequences of i.i.d. blocks.* [arXiv math.PR/1011.2679v2]

[31] I. D. Melamed. *Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons.* In Proceedings of the Third Workshop on Very Large Corpora, 1995.

[32] I. Dan Melamed. *Bitext maps and alignment via pattern recognition.* Comput. Linguist., **25**(1): 107–130, 1999.

[33] Mike Paterson and Vlado Dancik. *Longest common subsequences.* In Mathematical foundations of computer science 1994 (Kosice, 1994), volume 841 of Lecture Notes in Comput. Sci., pages 127–142. Springer, Berlin, 1994.

[34] Pavel Pevzner. *Computational Molecular Biology.* MIT Press. An algorithmic approach, A Bradford Book, Cambridge, 2000.

[35] Temple F. Smith and Michael S. Waterman. *Identification of common molecular subsequences.* J. Mol. Bio. **147**: 195–197, 1981.

[36] Michael J. Steele. *An Efron-Stein inequality for non-symmetric statistics* Annals of Statistics, **14**: 753–758, 1986.

[37] Felipe Torres. *On the probabilistic longest common subsequence problem for sequences of independent blocks.* PhD thesis, Bielefeld University, 2009. Online at http://bieson.ub.uni-bielefeld.de/volltexte/2009/1473/

[38] Michael S. Waterman *Estimating statistical significance of sequence alignments.* Phil. Trans. R. Soc. Lond. B, **344**(1): 383–390, 1994.

[39] Richard Arratia and Michael S. Waterman. *A phase transition for the score in matching random sequences allowing deletions.* Ann. Appl. Probab., **4**(1):200–225, 1994.

[40] Michael S. Waterman and M. Vingron. *Sequence comparison significance and Poisson approximation.* Statistical Science, **9**(3):367–381, 1994.

[41] Michael S. Waterman. *Introduction to Computational Biology.* Chapman & Hall, 1995.

[42] Kar Wing Li and Christopher C. Yang. *Automatic construction of english/chinese parallel corpora.* Journal of the American Society for Information Science and Technology **54**: 730–742, 2003.